(19)日本国特許庁（ＪＰ）　(12)公表特許公報（Ａ）　(11)特許出願公表番号
特表2001－511564
（P2001－511564A）
(43)公表日　平成13年８月14日(2001.8.14)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード(参考)
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３３０Ｃ	５Ｂ０７５
		15/40	３１０Ｆ	

審査請求　未請求　予備審査請求　有　（全 94 頁）

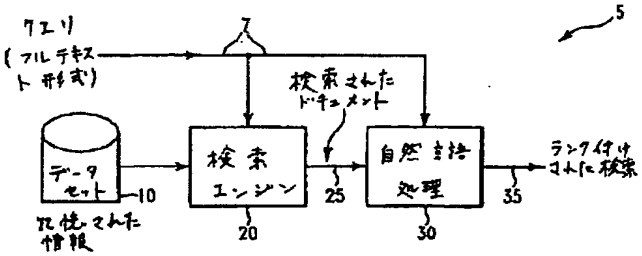
(21)出願番号　特願2000－504525(P2000－504525)
(86)(22)出願日　平成10年５月13日(1998.5.13)
(85)翻訳文提出日　平成12年１月24日(2000.1.24)
(86)国際出願番号　ＰＣＴ／ＵＳ９８／０９７１１
(87)国際公開番号　ＷＯ９９／０５６１８
(87)国際公開日　平成11年２月４日(1999.2.4)
(31)優先権主張番号　０８／８９８，６５２
(32)優先日　平成９年７月22日(1997.7.22)
(33)優先権主張国　米国（ＵＳ）
(81)指定国　ＥＰ(ＡＴ，ＢＥ，ＣＨ，ＣＹ，ＤＥ，ＤＫ，ＥＳ，ＦＩ，ＦＲ，ＧＢ，ＧＲ，ＩＥ，ＩＴ，ＬＵ，ＭＣ，ＮＬ，ＰＴ，ＳＥ)，ＣＮ，ＪＰ

(71)出願人　マイクロソフト　コーポレイション
Ｍｉｃｒｏｓｏｆｔ　Ｃｏｒｐｏｒａｔｉｏｎ
アメリカ合衆国　ワシントン州　98052　レドマンド　ワン　マイクロソフト　ウェイ（番地なし）
(72)発明者　プレイデン－ハーダー，リサ
アメリカ合衆国、20194　バージニア州、レストン、クリークベンド・ドライブ、12003
(74)代理人　弁理士　深見　久郎　（外５名）

最終頁に続く

(54)【発明の名称】　全体の精度を高めるためにサーチ結果の自然言語処理を用いる情報検索システムのための装置および方法

(57)【要約】
全体の精度を高めるために、たとえば従来の統計に基づくサーチエンジンのような情報検索エンジンによって検索された結果を処理するために自然言語処理を利用する情報検索システムのための装置およびそれに付随する方法を提供する。具体的には、このようなサーチは最終的に検索されたドキュメントの集合を生む。このような各ドキュメントは次に自然言語処理を受けて論理形式の集合を生じる。このような各論理形式は句内の語間の意味的関係、特に主題と修飾語句との構造を語－関係子－語の態様で符号化する。ユーザが与えるクエリも同様に分析されてそのための対応の論理形式の集合を生み出す。ドキュメントはドキュメントおよびクエリからの論理形式の予め規定された関数としてランク付けされる。具体的には、クエリのための論理形式の集合は、検索されたドキュメントの各々のための論理形式の集合と比較されて両方の集合内のこのような任意の論理形式間の一致を確認する。少なくとも１つの一致する論理形式を有する各ドキュメントがヒューリスティックにスコア付けされ、一致する論理形式のための異なる各関係が異なる対



【特許請求の範囲】

【請求項１】　記憶されているドキュメントをリポジトリから検索するための情報検索システムにおいて用いるための装置であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントを検索し、出力ドキュメント集合を規定するための検索システムを有し、前記装置は、

プロセッサと、

実行可能な命令が記憶されているメモリとを含み、

プロセッサはメモリに記憶されている命令に応答して、

クエリに応答してそのための第１の論理形式を生じ、第１の論理形式はクエリに関連した語の間の意味的関係を示し、

出力ドキュメント集合内のドキュメントの各別の１つに対して、対応する第２の論理形式を取得し、第２の論理形式は前記１つのドキュメント内の句に関連した語の間の意味的関係を示し、

クエリの第１の論理形式と、出力ドキュメント集合内の複数のドキュメントの各1つのための第２の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定し、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与える、装置。

【請求項２】　各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の１つのドキュメントに関連したレコードである、請求項１に記載の装置。

【請求項３】　クエリのための第１の論理形式と出力ドキュメント集合内の各別のドキュメントのための第２の論理形式との各々はそれぞれ、論理形式グラフ、そのサブグラフ、または論理形式三つ組のリストである、請求項２に記載の装置。

【請求項４】　プロセッサは記憶されている命令に応答して、

出力ドキュメント集合内のドキュメントの前記各別の1つのために、記憶媒体から対応の第２の論理形式を読出すか、または

出力ドキュメント集合内の前記各別の１つのドキュメントを分析することによ

って前記対応の第２の論理形式を生成する、請求項３に記載の装置。

【請求項５】　前記関数は、前記ドキュメントの１つのために、クエリに関連した前記第１の論理形式と前記１つのドキュメントに関連した前記第２の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、プロセッサは記憶されている命令に応答して、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定する、請求項４に記載の装置。

【請求項６】　クエリに関連した前記第１の論理形式または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記第２の論理形式は、それぞれ前記クエリにまたは前記ドキュメントの１つに関連した語句のパラフレーズをさらに含む、請求項５に記載の装置。

【請求項７】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の、対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項６に記載の装置。

【請求項８】　クエリに関連した前記第１の論理形式と出力ドキュメント集合内の任意のドキュメントに関連した前記第２の論理形式の任意のものとの間の前記一致は同一の一致である、請求項５に記載の装置。

【請求項９】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項８に記載の装置。

【請求項１０】　リポジトリはデータセットを含む、請求項５に記載の装置

。

【請求項１１】　クエリはフルテキストのクエリである、請求項５に記載の装置。

【請求項１２】　検索システムは統計的サーチエンジンを含む、請求項５に記載の装置。

【請求項１３】　ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライエントコンピュータからクエリを取得し、

出力ドキュメントの集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与える、請求項５に記載の装置。

【請求項１４】　前記サーバは複数の個別のサーバを含む、請求項１３に記載の装置。

【請求項１５】　検索システムは統計的サーチエンジンを含む、請求項１３に記載の装置。

【請求項１６】　ネットワーク接続はインターネットまたはインターネット接続である、請求項１５に記載の装置。

【請求項１７】　サーチエンジンはクエリに応答して、出力ドキュメントの集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項１６に記載の装置。

【請求項１８】　前記プロセッサおよび前記メモリを有するクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与える、請求項５に記載の装置。

【請求項１９】　検索システムは統計的サーチエンジンを含む、請求項１８に記載の装置。

【請求項２０】　ネットワーク接続はインターネットまたはインターネット接続である、請求項１９に記載の装置。

【請求項２１】　サーチエンジンはクエリに応答して、出力ドキュメントの集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項２０に記載の装置。

【請求項２２】　前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応じて前記検索システムを実現する、請求項５に記載の装置。

【請求項２３】　検索システムは統計的サーチエンジンを含む、請求項２２に記載の装置。

【請求項２４】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項５に記載の装置。

【請求項２５】　クエリはフルテキストのクエリである、請求項２４に記載の装置。

【請求項２６】　検索システムは統計的サーチエンジンを含む、請求項２４

に記載の装置。

【請求項２７】　ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライエントコンピュータからクエリを取得し、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与える、請求項２４に記載の装置。

【請求項２８】　サーバは複数の別個のサーバを含む、請求項２７に記載の装置。

【請求項２９】　検索システムは統計的サーチエンジンを含む、請求項２７に記載の装置。

【請求項３０】　ネットワーク接続はインターネットまたはインターネット接続である、請求項２９に記載の装置。

【請求項３１】　サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応じて、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項３０に記載の装置。

【請求項３２】　前記プロセッサおよび前記メモリを有するクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与える、請求項２４に記載の装置。

【請求項３３】　検索システムは統計的サーチエンジンを含む、請求項３２に記載の装置。

【請求項３４】　ネットワーク接続はインターネットまたはインターネット接続である、請求項３３に記載の装置。

【請求項３５】　サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項３４に記載の装置。

【請求項３６】　前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応答して前記検索システムを実現する、請求項２４に記載の装置。

【請求項３７】　検索システムは統計的サーチエンジンを含む、請求項３６に記載の装置。

【請求項３８】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項５に記載の装置。

【請求項３９】　クエリに関連した論理形式三つ組の前記第１のリストか、または出力ドキュメント集合内の前記ドキュメントの１つに関連した論理形式三つ組の前記第２のリストは、それぞれ前記クエリにまたは前記ドキュメントの１つに関連した語句のパラフレーズをさらに含む、請求項３８に記載の装置。

【請求項４０】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント

内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項３８に記載の装置。

【請求項４１】　前記関数は、クエリに関連した論理形式三つ組の少なくとも１つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的関係のタイプによって定義される、請求項３８に記載の装置。

【請求項４２】　プロセッサはメモリに記憶されている命令に応答して、

クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定し、

関連した少なくとも１つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各1つのために、前記一致する論理形式三つ組の各々に関連した意味的関係によって予め規定される重み数値を用いて前記各1つのドキュメント内の一致する論理形式三つ組に重み付けして、前記１つのドキュメントのための１つ以上の重みを形成し、

前記１つ以上の重みの関数として前記１つのドキュメントのためのスコアを計算し、

前記ドキュメントの各１つをその前記スコアに従ってランク付けしてランク順を規定する、請求項４１に記載の装置。

【請求項４３】　ランク順は重みの大きいものから小さいものの順である、請求項４２に記載の装置。

【請求項４４】　プロセッサはメモリに記憶されている命令に応答して、前記出力ドキュメント集合内のドキュメントの、最も高い、連続するランク付けを有する、前記出力ドキュメント集合のための前記エントリの第１の予め規定されたグループを提示する、請求項３８に記載の装置。

【請求項４５】　出力ドキュメント集合内の複数のドキュメントは、関連し

た少なくとも１つの一致する三つ組を有する、前記出力ドキュメント集合内のドキュメントからなる、請求項４４に記載の装置。

【請求項４６】　前記第１の論理形式三つ組および前記第２の論理形式三つ組の各々は、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項４５に記載の装置。

【請求項４７】　クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項３８に記載の装置。

【請求項４８】　クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項３８に記載の装置。

【請求項４９】　リポジトリはデータセットを含む、請求項３８に記載の装置。

【請求項５０】　クエリはフルテキストのクエリである、請求項３８に記載の装置。

【請求項５１】　検索システムは統計的サーチエンジンを含む、請求項３８に記載の装置。

【請求項５２】　ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライエントコンピュータからクエリを取得し、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与える、請求項３８に記載の装置。

【請求項５３】　サーバは複数の個別のサーバを含む、請求項５２に記載の装置。

【請求項５４】　検索システムは統計的サーチエンジンを含む、請求項５２に記載の装置。

【請求項５５】　ネットワーク接続はインターネットまたはインターネット接続である、請求項５４に記載の装置。

【請求項５６】　サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項５５に記載の装置。

【請求項５７】　前記プロセッサおよび前記メモリを有するクライエントコンピュータと、

ネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与える、請求項３８に記載の装置。

【請求項５８】　検索システムは統計的サーチエンジンを含む、請求項５７に記載の装置。

【請求項５９】　ネットワーク接続はインターネットまたはインターネット接続である、請求項５８に記載の装置。

【請求項６０】　サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力

ドキュメント集合内に含める、請求項５９に記載の装置。

【請求項６１】　前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応答して前記検索システムを実現する、請求項３８に記載の装置。

【請求項６２】　検索システムは統計的サーチエンジンを含む、請求項６１に記載の装置。

【請求項６３】　記憶されているドキュメントをリポジトリから検索するための情報検索システムにおいて用いるための方法であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントを検索し、出力ドキュメント集合を規定するための検索システムを有し、前記方法は、

クエリに応答してそのための第１の論理形式を生じるステップを含み、第１の論理形式はクエリに関連した語の間の意味的関係を示し、

出力ドキュメント集合内のドキュメントの各別の１つに対して、対応する第２の論理形式を取得するステップを含み、第２の論理形式は前記１つのドキュメント内の句に関連した語の間の意味的関係を示し、

クエリの第１の論理形式と、出力ドキュメント集合内の複数のドキュメントの各１つのための第２の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定するステップと、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与えるステップとを含む、方法。

【請求項６４】　各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の１つのドキュメントに関連したレコードである、請求項６３に記載の方法。

【請求項６５】　クエリのための第１の論理形式と出力ドキュメント集合内の各別のドキュメントのための第２の論理形式との各々はそれぞれ、論理形式グラフ、そのサブグラフ、または論理形式三つ組のリストである、請求項６４に記載の方法。

【請求項６６】　前記取得するステップは、

出力ドキュメント集合内のドキュメントの前記各別の１つのために、記憶媒体から対応の第２の論理形式を読出すか、または

出力ドキュメント集合内の前記各別の１つのドキュメントを分析することによって、前記対応の第２の論理形式を生成するステップを含む、請求項６５に記載の方法。

【請求項６７】　前記関数は、前記ドキュメントの１つのために、クエリに関連した前記第１の論理形式と前記１つのドキュメントに関連した前記第２の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、前記ランク付けするステップは、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定するステップを含む、請求項６６に記載の方法。

【請求項６８】　クエリに関連した前記第１の論理形式または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記第２の論理形式は、それぞれ前記クエリにまたは前記ドキュメントの１つに関連した語句のパラフレーズをさらに含む、請求項６７に記載の方法。

【請求項６９】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の、対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項６８に記載の方法。

【請求項７０】　クエリに関連した前記第１の論理形式の任意のものと出力ドキュメント集合内の任意のドキュメントに関連した前記第２の論理形式の任意のものとの間の前記一致は同一の一致である、請求項６７に記載の方法。

【請求項７１】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の、対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、

対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項７０に記載の方法。

【請求項７２】　リポジトリはデータセットを含む、請求項６７に記載の方法。

【請求項７３】　クエリはフルテキストのクエリである、請求項６７に記載の方法。

【請求項７４】　検索システムは統計的サーチエンジンを含む、請求項６７に記載の方法。

【請求項７５】　システムはクライエントコンピュータをさらに含み、前記方法はクライエントコンピュータにおいて、

ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとを含み、

システムはネットワーク接続を介してクライエントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライエントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与えるステップとを含む、請求項６７に記載の方法。

【請求項７６】　検索システムは統計的サーチエンジンである、請求項７５に記載の方法。

【請求項７７】　ネットワーク接続はインターネットまたはインターネット接続である、請求項７６に記載の方法。

【請求項７８】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出される得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、

出力ドキュメント集合内に含めるステップをさらに含む、請求項７７に記載の方法。

【請求項７９】　システムはクライエントコンピュータとネットワーク接続を介してクライエントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与えるステップをさらに含む、請求項６７に記載の方法。

【請求項８０】　検索システムは統計的サーチエンジンを含む、請求項７９に記載の方法。

【請求項８１】　ネットワーク接続はインターネットまたはインターネット接続である、請求項８０に記載の方法。

【請求項８２】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメントの集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、クライエントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項８１に記載の方法。

【請求項８３】　システムはコンピュータをさらに含み、前記方法はコンピュータにおいて前記検索システムを実現するステップを含む、請求項６７に記載の方法。

【請求項８４】　検索システムは統計的サーチエンジンを含む、請求項８３に記載の方法。

【請求項８５】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特

定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項６７に記載の方法。

【請求項８６】　リポジトリはデータセットを含む、請求項８５に記載の方法。

【請求項８７】　クエリはフルテキストのクエリである、請求項８５に記載の方法。

【請求項８８】　検索システムは統計的サーチエンジンを含む、請求項８５に記載の方法。

【請求項８９】　システムはクライエントコンピュータをさらに含み、前記方法はクライエントコンピュータにおいて、

ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとをさらに含み、

システムはネットワーク接続を介してクライエントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライエントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与えるステップとをさらに含む、請求項８５に記載の方法。

【請求項９０】　検索システムは統計的サーチエンジンを含む、請求項８９に記載の方法。

【請求項９１】　ネットワーク接続はインターネットまたはインターネット接続である、請求項９０に記載の方法。

【請求項９２】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項９１に記載の方法

。

【請求項９３】　システムはクライエントコンピュータとネットワーク接続を介してクライエントコンピュータに接続されるサーバとを含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与えるステップをさらに含む、請求項８５に記載の方法。

【請求項９４】　検索システムは統計的サーチエンジンを含む、請求項９３に記載の方法。

【請求項９５】　ネットワーク接続はインターネットまたはインターネット接続である、請求項９４に記載の方法。

【請求項９６】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、クライエントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項９５に記載の方法。

【請求項９７】　システムはコンピュータをさらに含み、前記方法はコンピュータにおいて前記検索システムを実現するステップを含む、請求項８５に記載の方法。

【請求項９８】　検索システムは統計的サーチエンジンを含む、請求項９７に記載の方法。

【請求項９９】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項６７に記

載の方法。

【請求項１００】　クエリに関連した論理形式三つ組の前記第１のリストか、または出力ドキュメント集合内の前記ドキュメントの１つに関連した論理形式三つ組の前記第２のリストは、それぞれ前記クエリにまたは前記ドキュメントの１つに関連した語句のパラフレーズをさらに含む、請求項９９に記載の方法。

【請求項１０１】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項９９に記載の方法。

【請求項１０２】　前記関数は、クエリに関連した論理形式三つ組の少なくとも１つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的関係のタイプによって定義される、請求項９９に記載の方法。

【請求項１０３】　前記ランク付けするステップは、

クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定するステップと、

関連した少なくとも１つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各1つのために、前記一致する論理形式三つ組の各々に関連した意味的関係によって予め規定される重み数値を用いて前記各1つのドキュメント内の一致する論理形式三つ組に重み付けして、前記１つのドキュメントのための１つ以上の重みを形成するステップと、

前記１つ以上の重みの関数として前記１つのドキュメントのためのスコアを計算するステップと、

前記ドキュメントの各１つをその前記スコアに従ってランク付けしてランク順

を規定するステップとを含む、請求項１０２に記載の方法。

【請求項１０４】　ランク順は重みの大きいものから小さいものの順である、請求項１０３に記載の方法。

【請求項１０５】　記憶されているエントリを与えるステップは、前記出力ドキュメント集合内のドキュメントの、最も高い、連続するランク付けを有する、前記出力ドキュメント集合のための前記エントリの第１の予め規定されたグループを提示するステップを含む、請求項９９に記載の方法。

【請求項１０６】　出力ドキュメント集合内の前記複数のドキュメントは、関連した少なくとも１つの一致する三つ組を有する、前記出力ドキュメント集合内のドキュメントからなる、請求項１０５に記載の方法。

【請求項１０７】　前記第１の論理形式三つ組および前記第２の論理形式三つ組の各々は、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項１０６に記載の方法。

【請求項１０８】　クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項９９に記載の方法。

【請求項１０９】　クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項９９に記載の方法。

【請求項１１０】　リポジトリはデータセットを含む、請求項９９に記載の方法。

【請求項１１１】　クエリはフルテキストのクエリである、請求項９９に記載の方法。

【請求項１１２】　検索システムは統計的サーチエンジンを含む、請求項９９に記載の方法。

【請求項１１３】　前記システムはクライエントコンピュータをさらに含み

、前記方法はクライエントコンピュータにおいて、

ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとを含み、

前記システムはネットワーク接続を介してクライエントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライエントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与えるステップとをさらに含む、請求項９９に記載の方法。

【請求項１１４】　検索システムは統計的サーチエンジンを含む、請求項１１３に記載の方法。

【請求項１１５】　ネットワーク接続はインターネットまたはインターネット接続である、請求項１１４に記載の方法。

【請求項１１６】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連するサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項１１５に記載の方法。

【請求項１１７】　システムはクライエントコンピュータとネットワーク接続を介してクライエントコンピュータに接続されるサーバとを含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライエントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライエントコンピュータに与えるステップをさらに含む、請求項９９に記載の方法。

【請求項１１８】　検索システムは統計的サーチエンジンを含む、請求項１１７に記載の方法。

【請求項１１９】　ネットワーク接続はインターネットまたはインターネット接続である、請求項１１８に記載の方法。

【請求項１２０】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、クライエントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項１１９に記載の方法。

【請求項１２１】　システムはコンピュータをさらに含み、前記方法はコンピュータにおいて前記検索システムを実現するステップをさらに含む、請求項９９に記載の方法。

【請求項１２２】　検索システムは統計的サーチエンジンを含む、請求項１２１に記載の方法。

【請求項１２３】　コンピュータで実行可能な命令を記憶し、請求項６３に記載のステップを実行するためのコンピュータ読出可能媒体。

【発明の詳細な説明】

【０００１】

【発明の分野】

本発明は、たとえば統計に基づいた従来のサーチエンジンなどの情報検索エンジンによって検索された結果を処理するために自然言語処理を利用して全体的な精度を改善する、情報検索システムのための装置およびそれに付随する方法に関する。

【０００２】

【先行技術の説明】

数十年から現在に至るまで、自動情報検索技術は、出版物および／またはそのための書誌情報を含む従来のデータベースなどの大容量データ記憶装置から記憶情報を検索するためにますます頻繁に使用されている。このような従来のデータベースはたとえば米国電気電子通信学会（ＩＥＥＥ）によって維持されており、かつたとえばKnight-Ridder Information Inc.のダイアログ（Dialog）情報サービスによって現在アクセス可能である、ＩＮＳＰＥＣデータベースの場合のように、電気工学およびコンピュータ関連技術などの、広い範囲ではあるが特定のトピックに向けられる情報を一般的に含むため、専門的になる傾向がある（DIALOGはKnight-Ridder Information Inc.の登録サービスマークである）。このタイプのデータベースは明らかに、関連の論文および他の出版物の出版数の増加に伴い増大し続けることは確実だが、この増大は比較的なだらかで、適度にうまく調整される傾向がある。さらに、このように専門化されたデータベースは比較的うまく組織化されやすい。

【０００３】

しかしながら、インターネットを介してアクセス可能ないわゆる「ワールドワイドウェブ」（以下単に「ウェブ」と称する）の出現および隆盛により、また従来の出版とは対照的にウェブに情報を投稿しかつそれから情報にアクセスする、比較的容易で低コストで行えることにより、ウェブ上で利用できる情報量は、爆発的ではないにしても非常に指数関数的に近い増加を遂げ、現実的な視界内に制限があるようには思われない。ウェブはおよそ人間が試みる全ての学問分野に及

ぶ、増々膨大な量の情報を提供するが、ウェブ上の情報内容は非常に無秩序であり、極端な程に組織化されておらず、そのためウェブからの情報のアクセスおよび検索を非常に複雑にし、多大な労苦を強いる。

【０００４】

ウェブからの情報の検索作業をはるかに容易にするために、過去数年にコンピュータによる多数のサーチエンジンが開発されており、広く一般的に使用されている。概して、これらの従来のエンジンは、ソフトウェアにより実装された「ウェブクローラ」によって、ウェブサイトに自動的に入り、その中のハイパーテキストリンクを順次追跡し、その中にある各ドキュメントを抽出し、要約し、かつそれにいわゆる「キーワード」によって索引を付けて大型データベースとし、後にアクセスできるようにする。具体的には、このような要約により、クローラが遭遇するこのような各ドキュメントは通常「１袋の単語」と呼ばれるものにまで凝縮される。これは、意味論上および統語論上の全ての情報は除去されているが、ドキュメント内に存在する内容語を含む。内容語は、ドキュメント自体にあることもあるし、および／またはそのドキュメントのハイパーテキストマークアップ言語（ＨＴＭＬ）版の記述フィールドにだけ現われることもある。いずれにせよ、エンジンは、そうしたドキュメントに対するエントリ、すなわちドキュメントレコードを作成する。各ドキュメントについて、その内容語の各々に索引が付けられて、そのドキュメントに戻るリンクを有する、サーチ可能なデータ構造が形成される。ドキュメントレコードは典型的には、（ａ）ウェブアドレス、すなわちＵＲＬ－－これによってウェブブラウザによって対応のドキュメントをアクセスすることができるユニフォームリソースロケータと、（ｂ）そのドキュメント内の、さまざまな内容語とを含み、さらに、エンジンによっては、そのドキュメント内の他の内容語に対する、これら内容語の各々の相対的なアドレスを含み、さらに（ｃ）概要とを含み、これはドキュメントのうち数行のみであるか、またはドキュメントの最初の数行であることが多いが、さらに場合によってはｄ）そのＨＴＭＬの記述フィールド内に記載された、そのドキュメントに関する説明とを含む。データベースをサーチするために、ユーザはキーワードに基づいたクエリをエンジンに与える。

【０００５】

クエリは典型的には、ユーザによって与えられた１または２以上、多くの場合には小さな数だけのキーワードを含み、エンジンの能力に依存して、場合によっては連続したキーワード間にあるブール論理（たとえば「ＡＮＤ」または「ＯＲ」）、または類似したオペレータ（たとえば数的近さ）を含む。クエリに応答して、エンジンはできるだけ多くのキーワードと、論理的なまたは近さに関するオペレータが提供されている場合には、リクエストされた特定の組合わせまたはお互いにある「レンジ」（特定の数の、内容語）内にあるようなキーワードとを含むドキュメントを突き止めようとする。これを行なう際に、エンジンはそのデータベースをサーチして、クエリのキーワードの１つと一致する少なくとも１つの単語を含み、リクエストがあった場合にはそのリクエストによって特定されたオペレータおよび／またはレンジに一致するような、ドキュメントを突き止める。エンジンは見つけ出したこのようなドキュメントの各々に対して、それに関するドキュメントレコードを検索し、そのドキュメント内のキーワード一致の数に従って、同様の他のこうした文書に対するランク付けをしてそのレコードをユーザに提示する。

【０００６】

ユーザによって与えられるキーワードのクエリに応答して検索されただけのドキュメントの大部分はクエリとは全く関連がないことが多く、ユーザをいらだたせる。

【０００７】

したがって、無関係なドキュメントが検索される数を減らすために、キーワードに基づいた従来のサーチエンジン（以下、単に「統計的サーチエンジン」と称する）は、統計的処理をそれらのサーチ法に取り入れている。たとえば、クエリ内のキーワードと検索された各ドキュメントレコード中の内容語との間で一致するキーワードの総数およびこれらの単語の一致の程度、すなわち組合せとしておよび／またはリクエストされた近さのレンジ内にあるか否かに基づいて、統計的サーチエンジンは、検索されたこのようなドキュメントレコードの各々に対して、「統計値」と包括的に呼ばれることが多い数値尺度を計算する。これらの統計

値は、一致する各単語に対するドキュメント頻度の逆数を含み得る。その後エンジンはそれらの統計値によってドキュメントレコードをランク付けし、最もランクの高い、典型的には５個から２０個以下の、予め規定された少数の検索済レコードに関するドキュメントレコードを、ユーザに戻す。検索された第１のグループのドキュメントに対するユーザが第１のグループのドキュメントレコード（またはある種のエンジンのように、エンジンによってドキュメントが戻される場合にはドキュメント自体）をユーザが検討すると、ユーザは、次に高いランキングのドキュメントレコードのグループを要求することができ、以下同様に検索されたドキュメントレコード全てがこうして検討されるまで、要求をすることができる。

【０００８】

従来、サーチエンジンの性能は再現および精度によって評価されてきた。再現は、データセット内の関連の全てのドキュメントに対し、所与のクエリに応答して実際に検索されたこのようなドキュメントの数を百分率で測る。一方、精度は、検索された全てのドキュメントに対し、クエリに実際に関連するドキュメントの数を百分率で測る。最終的に検索されるドキュメントの単なる数は重要ではないため、ウェブサーチエンジンについて考える場合には、再現は性能に関する重要な測定基準ではないと考える。実際に、クエリによってはこの数は過度に大きいこともある。したがって、有用な結果を生み出すためには、エンジンによって索引付けされた関連のドキュメントの全てを取り出す必要はないと考える。しかしながら精度は極めて重要であると考えられる。すなわち、ランクが最も高く最初にユーザに提示されるドキュメントは、クエリに最も関連するものであるべきであると考える。

【０００９】

従来の統計的サーチエンジンの精度が比較的低いのは、単語が独立変数である、すなわち全ての文章の単語は互いに独立して現われるという仮定に基づいていることに由来する。この場合の独立とは、別のある単語が文書中に存在するときにそのドキュメントに任意の１つの単語が表れるという条件付確率が常にゼロであること、すなわち、ドキュメントが、構造を持たない単語の集まりを含むだけ

であるか、または単に「１袋の単語」でしかないことを意味する。容易に認識できるように、この仮定は全ての言語に関して非常に誤ったものである。英語は、他の言語と同様に、単語に関する、膨大な量で複雑な統語論上および語彙－意味論上の構造を有し、これらの単語の意味は、使用される特定の言語的文脈に基づいて、しばしば広く、異なることが多く、文脈はその場合にも単語に与えられた意味と、いかなる単語が後に現われるかとを決定する。したがって、文章に表れる単語は単に独立しているのでは全くなく、相互に高度に依存する。キーワードに基づいたサーチエンジンはこのきめの細かな言語的構造を全く無視している。たとえば、自然言語で表現された「How many hearts does an octopus have?」というクエリの例について考える。内容語「hearts」および「octopus」またはその形態素的語幹に基づいて動作する統計的サーチエンジンであれば、その材料の部分に、したがってその内容を表わす単語として「artichoke hearts(アーチチョークの芯)からsquid（イカ）、onions（タマネギ）およびoctopus（タコ）」を有するレシピを含む、記憶されたドキュメントを戻すか、またはユーザをそのドキュメントに導く。内容を表わす２つの内容単語「octopus」および「hearts」が一致するので、このエンジンは、たとえば近さおよび論理的なオペレータを含む統計的測定値に基づいて、実際にはドキュメントがクエリとはかなり無関係であっても、このドキュメントが優れた一致であると決定してしまう。

【００１０】

この技術分野では、ラベル付きでない関係にあるヘッド－修正語対として統語論的な句の要素を抽出するためのさまざまな方策が教示されている。これらの要素はその後、従来の統計的ベクトル－空間モデルにおける（典型的には内部構造のない）述語として索引付けされる。

【００１１】

このような方策の一例はJ. L. Fagan,"Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods", Ph.D. Thesis, Cornell University, 1988, pages i-261に記載されている。具体的には、この方策は自然言語処理を使用して英語の文章を分析して、統語論的な句の構成要素を抽出し、これらの句の構成要素は後に術語として

扱われ、統計的なベクトルー空間モデルを用いた索引に索引付けされる。検索時に、はユーザは自然言語でクエリを入力し、この方策の下では、このクエリに自然言語処理が施され、分析され、索引中検索にレコードされている憶された要素に類似した統語論的な句の構成要素の要素を抽出する。その後、クエリによる統語論的な句の構成要素と、索引に記憶されているたものとのを比較照合一致さが行なわれるせる試みがなされる。著者は、この純粋に統語論的な方策を、統語論的な句の構成要素を特定するために確率論的方法を使用すると統計的な方策とを対比させており、この統計的方策では統語論的な句の構成要素を特定するために確率論的方法が使用されるいる。著者は、自然言語処理でははは確率論的方策からをの大幅な実質的に改善は見られずせず、自然言語処理が時に提供し得る精度がのわずかに軽微な改善されることによっては、自然言語処理に伴うかなりの処理コストはを正当化するものではできない、と結論し付けている。

【００１２】

サーチ用のクエリに含ませるための適切な単語を選択するための自然言語処理を使用する場合の、このような統語論に基づいた別の方策が、T. Strzalkowski, “Natural Language Information Retrieval: TIPSTER-2 Final Report”, Proceedings of Advances in Text Processing: Tipster Program Phase 2, DARPA, 6-8 May 1996, Tysons Corner, Virginia, pages 143-148（以下、「DARPAの論文」と呼ぶ）およびT. Strzalkowski, “Natural Language Information Retrieval”, Information Processing and Management, Vol. 31, No. 3, 1995, pages 397-417に記載されている。この方策は理論的には可能性を秘めたものではるが、ＤＡＲＰＡの論文の第１４７頁から１４８頁において著者は、基盤となる自然言語技術を実現するために要求される処理が高度であるために、この方策は現在のところは実用的でない、と結論している。

「...ただし、我々の性能要件を満たす（またはこれらの要件に少なくとも近いと考えられる）ＮＬＰ［自然言語処理］技術は、依然として、自然言語の文を扱う能力において性能がかなり低い。特に、概念的構成および論理形式などにかかわる進歩した処理には、計算の面から依然として届かない。これらの進歩した技術は表現レベルの限界の問題に対処するものであるため、より有効となるであろ

うと仮定することもできるが、実証に乏しく、しかもかなり小さなスケールのテストに限定されざるを得ない。」

この種の統合論に基づいたさらなる方策がB. Katz, "Annotating the World Wide Web Using Natural Language", Conference Proceedings of RIAO 97, Computer-Assisted Information Searching in Internet, McGill University, Quebec, Canada, 25-27 June 1997, Vol. 1, pages 136-155（以下、「Katzの出版物」と称する）に記載されている。Katzの出版物の記載によれば、内部構造を維持したままで主語－動詞－目的語表現が作成されるので、検索時には軽微な統語論上の変更に対処できる。

【００１３】

これらの統語論的方策が大した改善をもたらさなかったこと、またはその時点で利用可能な自然言語処理システムによっては実現できなかったことにより、研究分野は、クエリによる最初の結果の精度および再現を直接改善しようとする試みから、ユーザインターフェイスの改善へ、すなわち具体的には、「類語を検索せよ」とユーザが検索結果に応答することなどの、ユーザとの相互作用に基づいてクエリの精度を向上させる方法、および適当な固まりに分けて結果を表示することを含む、クエリに対する結果を視覚化するための方法による改善に、移行した。

【００１４】

これらの改善自体は有用であるが、これらの改善によって達成可能な精度の向上は依然として失望するほど少なく、キーワードによるサーチに特有なユーザの感ずる歯がゆさを大幅に軽減するには明らかに不十分である。具体的には、ユーザには、関連ある応答がまばらにしかないような比較的大きなドキュメントの集合を手操作によってふるいにかけることが依然として要求される。

【００１５】

したがって、この技術分野においては、情報検索に対する従来の統計的方策によって達成可能なものに優る著しい精度の改善をもたらすことができる、情報を検索するための技術、特定的には装置およびそれに付随する方法が必要とされている。さらに、このような技術は、任意に生ずる文章内の、広範囲な文のタイプ

および長さに対しても信頼性が高くかつ反復可能な結果をもたらし、かつ実用的であり実現の際のコスト面でも有効である必要がある。このような従来の方策の精度に対して、しかもこの技術分野に特有な問題にもかかわらず、精度を著しく改善するためには、このような技術は好ましくは、意味的な内容とクエリの内容との照合に基づいて、関連のドキュメントを選択し検索してその後ユーザに提示するという効果を得るために自然言語処理を利用するべきである。

【００１６】

【発明の概要】

我々の広い教示によると、本発明は、たとえば統計的ウェブサーチエンジンによって行なわれるキーワードに基づいたドキュメントサーチの精度を改善するために自然言語処理を採用することにより、この必要を満たす。

【００１７】

大まかに言って、この処理は、それぞれがサーチクエリおよび検索されたドキュメントの各々に関連した論理形式を生成し、比較し、それらの一致を重み付けすることを含む。検索されたドキュメントは、クエリおよび検索されたドキュメントの両方に関する「論理形式」の予め規定された関数に基づいて、特定的には、ドキュメントに関連した一致する論理形式に関連する重みの和に基づいて、ランク付けされ、その順で最終的に表示される。論理形式とは、任意のサイズのテキストを表わす単語がラベル付きの関係によってリンクされる有向非循環グラフである。特に、論理形式は入力文字列における重要な単語間の意味論上の関係を、特に主題および修飾語句の関係を描く。この描写はさまざまな特定の形式をとることができ、たとえば論理形式グラフまたは、たとえば論理形式三つ組（triple）のリストを含むその任意のサブグラフの形式をとることができる。ここで、三つ組の各々はたとえば「語－関係詞－語」という形式をとるが、これら形式のいずれでも我々の発明に用いることができる。

【００１８】

我々の特定の教示によると、このようなサーチにより最終的には、たとえばデータベースまたはワールドワイドウェブからの検索されたドキュメントの組が得られる。その後各ドキュメントには、自然言語処理が施され、特定的には形態素

的、統語論的および論理的な形式に関する処理が施され、最終的には各ドキュメントの各文に対して適当な論理形式が生成される。ユーザによって与えられたクエリが同様に分析され、それに関する対応の論理形式三つ組の集合が得られる。クエリに対する論理形式の組はその後、検索されたドキュメントの各々に関連した論理形式の組と比較され、クエリの組からの論理形式と各ドキュメントの組からの論理形式との一致が確認される。一致を生じないドキュメントはそれ以上は考慮されない。残りの各ドキュメントはその後ヒューリスティックにスコア付けされる。特に、種々のタイプの関係の各々、すなわち論理形式内に現われ得る深層主語、深層目的語および機能語などに、予め規定された重みが割当てられる。このような残りのドキュメントの各々のスコアは、その中にある一致する論理形式の重みの、予め規定された関数である。この関数は、例えば、そのドキュメントに現われる、一意な、一致する三つ組（二重の一致を無視）全てに関連した重みの和でもよい。最後に、保持されたドキュメントが、それらのスコアに基づいて、ユーザの選択にしたがい降順に、典型的にはたとえば５個または１０個というような予め規定された少数のグループで、最も高いスコアを有するグループから始まり、順に他のグループという順番でユーザに提示される。

【００１９】

本発明は種々のいくつかの処理トポロジーで使用することができる。すなわち、（ａ）クエリおよびキーワードに基づいたサーチ（ドキュメント検索）の両方がローカルなパーソナルコンピュータ（ＰＣ）などの共通のコンピュータによって処理される場合、（ｂ）キーワードに基づいたサーチがたとえばリモートサーバであるリモートコンピュータによって処理され、クエリおよびサーチ結果がたとえばクライアントＰＣによって処理される場合、または（ｃ）クエリがクライアントＰＣで作成され、残りの処理がさまざまなリモートサーバに分配される場合、である。さらに、データベースの各ドキュメントを索引付けしてデータベース化する際に前処理して、関連の論理形式を得て、これらの論理形式を記憶しておいて後にアクセスできるようにすることによって、そのドキュメントが後に検索されて自然言語処理を受ける場合には常に実行時間が節約されるようになる。

【００２０】

【詳細な説明】

本発明の教示は添付の図面を参照して以下の詳細な説明を考慮すると容易に理解できる。

【００２１】

理解を容易にする目的で、可能な場合には、図面に共通する要素には同一の参照番号を付す。

【００２２】

以下の説明を考慮すれば、当業者であれば、サーチエンジンが従来の統計的エンジンであるか否かに関係なく、我々の本発明の教示をほとんど全ての情報検索システムに容易に利用して、そのシステムで使用されるサーチエンジンの精度を高めるようにできることが明らかに認識できるであろう。さらに、我々の発明は、磁気媒体、光媒体（たとえばＣＤ－ＲＯＭ）または他の媒体に記憶されているデータベースなどのほぼ全てのタイプの大容量記憶装置からテキスト形式の情報を検索する際に、たとえば英語、スペイン語およびドイツ語など、テキスト形式の情報がどの言語であるかにかかわらず、精度を改善するために利用できる。

【００２３】

広く言えば、我々の本発明によると、たとえば検索エンジンで使用されているサーチエンジンによって提供されるレコードを、たとえば究極的にはドキュメントを、フィルタリングしランク付けするために自然言語処理を用いることによって、検索エンジンの精度を著しく向上させることができることが我々には認識できた。

【００２４】

この点に留意して、図１は我々の発明を利用する情報検索システム５の非常に高いレベルのブロック図を示す。システム５はたとえばキーワードに基づいた統計的検索エンジンである従来の検索エンジン２０と、その後に続くプロセッサ３０とを含む。プロセッサ３０は、後に説明するように、我々の発明の自然言語処理技術を利用して、エンジン２０によって生成されたドキュメントをフィルタリングして再度ランク付けし、ユーザによって与えられたクエリに対する関連性がこの技術を仕様しない場合よりも高い、順序付けのされた検索されたドキュメン

トの集合をもたらす。

【００２５】

具体的には、動作時にユーザはサーチ用のクエリをシステム５に与える。クエリには、自然言語処理によってその意味論上の内容を最大限に利用し、それによってエンジン２０だけにより得られる精度よりもさらに精度を向上させるために、（通常は「リテラル」と呼ばれる）フルテキスト形式が用いられる。システム５はこのクエリをエンジン２０およびプロセッサ３０の両方に与える。クエリに応答して、エンジン２０は、記憶されたドキュメントのデータセット１０をサーチし、それから検索されたドキュメントの集合を出力する。このドキュメントの集合（「出力ドキュメント集合」とも呼ばれる。）は、線２５で表わされるように、入力としてプロセッサ３０に与えられる。後に詳細に説明するように、プロセッサ３０内では、集合内のドキュメントの各々に対して、自然言語処理、特に形態素的、統語論的および論理的な形式に関する処理が施され、そのドキュメント内の各文に対する論理形式を生成する。ある文に関するこのような論理形式の各々は、その文内の言語学的な句の中の単語間の意味論的な関係、特に主題および修飾語句構造を符号化したものである。プロセッサ３０は同一の態様でクエリを分析し、それに関する対応の論理形式の集合を生成する。そしてプロセッサ３０は、クエリに関する形式の集合と、その集合内のドキュメントの各々に関連する論理形式の集合とを比較し、クエリ集合内の論理形式と各ドキュメントに関する論理形式との間に一致があるかどうかを確認する。一致を生まないドキュメントはそれ以上考慮されない。残りの、クエリに関する論理形式と一致する少なくとも１つの論理形式を含むドキュメントの各々は、プロセッサ３０によって保持され、ヒューリスティックにスコア付けされる。後に説明するように、異なった各タイプの関係、すなわち、論理形式三つ組に現われ得る深層主語、深層目的語および機能語などに対して、予め規定された重みが割当てられる。このようなドキュメントの各々の合計の重み（すなわちスコア）は、たとえば一致する一意な三つ組、すなわち二重に一致する三つ組を無視したものの全ての重みの和である。最後に、プロセッサ３０は、保持されたドキュメントを、それらのスコアに基づいてランク付けして、たとえば５個とか１０個とかという予め定められた数の

グループに分けて、スコアの最も高いものからユーザに提示する。

【００２６】

システム５が非常に汎用的であり、広範囲な種々のアプリケーションに適合させることができるので、以下の議論を簡単にするために、我々は１つの例を用いて我々の発明の用途を議論することとする。この例は、ワールドワイドウェブからのドキュメントであって、索引付けされたデータセットを形成する英語のドキュメントの、格納されたレコードを検索するために、従来のキーワードに基づいた統計的インターネットサーチエンジンを採用する情報検索システムであろう。このような各レコードは一般に、以下に説明するように、対応のドキュメントに関する予め規定された情報を含む。他のサーチエンジンの場合、レコードがドキュメント自体の全体を含んでもよい。以下の議論では、ドキュメントに関する、そのドキュメントを見出すことができるウェブアドレスを含むある情報を含むレコードを検索する従来のインターネットサーチエンジンに使用する場合を例として我々の発明を扱うが、包括的に言えば、そのエンジンによって検索される究極的な項目とは、ウェブからドキュメントに実際にアクセスするために一般的にはそのアドレスを用いる中間的な処理が採用されるとしても、実際にはドキュメントである。以下の説明を考慮すると、我々の本発明が、他のいかなる情報検索の適用例の使用にも容易に適合可能であることが当業者には容易に認められるであろう。

【００２７】

図２は、インターネットサーチエンジンの例において使用されている、我々の発明の特定の実施例の高いレベルでのブロック図を示す。我々の発明は、主にこの特定の実施例を例として詳細に説明する。図示されるように、システム２００はクライアントパーソナルコンピュータ（ＰＣ）などのコンピュータシステム３００を含み、これは、ネットワーク接続２０５を介して、ネットワーク２１０（ここではインターネットであるが、たとえばイントラネットなどの他のこのようなネットワークをこれに代えて用いてもよい）およびネットワーク接続２１５によってサーバ２２０に接続される。サーバは典型的にはコンピュータ２２２を含み、これはインターネットサーチエンジン２２５をホストし、たとえばＡＬＴＡ

ＶＩＳＴＡサーチエンジン（ＡＬＴＡ　ＶＩＳＴＡはマサチューセッツ州メイナード（Maynard, Massachusetts）のDigital Equipment Corporationの登録商標である。）が典型であり、大容量データ記憶装置２２７に接続され、これは典型的には、サーチエンジンによって索引づけられ、インターネット上のワールドワイドウェブによってアクセス可能であるドキュメントレコードのデータセットである。このようなレコードの各々は典型的には、（ａ）ウェブブラウザによって対応のドキュメントがアクセス可能であるウェブアドレス（通常はユニフォームリソースロケーターＵＲＬと呼ばれる）と、（ｂ）そのドキュメントに現われる予め規定された内容語とを含み、エンジンによっては、そのドキュメント内の、他の内容語に対するこのような各単語の相対的なアドレスを含み、さらに（ｃ）ドキュメントのうち数行だけであるか、またはドキュメントの最初の数行であることが多い概要と、（ｄ）ハイパーテキストマークアップ言語（ＨＴＭＬ）記述フィールドに提供されるような、ドキュメントの説明とを含む。

【００２８】

コンピュータシステム３００に配置されたユーザは、このシステムで動作する、たとえば関連のウェブブラウザ（たとえばMicrosoft Corporationから入手可能であり、我々の発明の教示を含むよう適当に変形された「インターネットエクスプローラ」バージョン３．０に基づくもの）を介して、サーバ２２０、特にそこで動作するサーチエンジン２２２へのインターネット接続を確立する。さらにユーザは、ここでは線２０１によって表わされるクエリをブラウザに入力し、ブラウザはシステム３００を介して、サーバ２２０へのインターネット接続によって、サーチエンジン２２５にクエリを送信する。するとサーチエンジンはデータセット２２７に記憶されたドキュメントレコードに対してクエリを処理し、エンジンがクエリに関連すると判断したドキュメントに対する検索レコードの集合を生成する。エンジン２２５がどのようにしてドキュメントを索引付けしドキュメントレコードを形成してデータ記憶装置２２７に記憶するか、および記憶されたこのようなドキュメントを選択するためにエンジンが実際にはどのような分析を行なうかはいずれも本発明と無関係であるため、これらの局面はいずれもこれ以上詳細には説明しない。クエリに応答して、エンジン２２５が、検索されたドキ

ュメントレコードの集合をインターネット接続を介してウェブブラウザ４２０に返す、といえば十分である。ブラウザ４２０は、エンジン２２５がドキュメントを検索するのと同時に、かつ／またはその後に、クエリを分析して、その、論理形式三つ組の対応の集合を生成する。サーチエンジンがそのサーチを完了し、ドキュメントレコードの集合を取出し、その集合をブラウザに与えると、対応のドキュメント（すなわち出力ドキュメントの集合を形成するもの）自体が関連のウェブサーバからブラウザによってアクセスされる（これに関連したデータセットは全体として保存されたドキュメントの「リポジトリ」を形成する。このようなリポジトリは、たとえば独立したＣＤ－ＲＯＭベースのデータ検索アプリケーションなどにおけるもののように、スタンドアローンのデータセットであってもよい）。するとブラウザはアクセスされたドキュメント（すなわち出力ドキュメントの内のもの）の各々を分析し、このようなドキュメントの各々に関する、論理形式三つ組の、対応する集合を形成する。その後、後に詳細に説明するように、ブラウザ４２０は、クエリと、検索されたドキュメントとの間の論理形式三つ組の一致照合に基づいて、このような一致を有する各ドキュメントをスコア付けし、それらのドキュメントを、線２０３によって表わされるようにランクの降順に、ブラウザを通じたユーザの選択にしたがって、典型的には最も高いランクを有する少数の予め規定されたドキュメントのグループとして、大きなスコアのものからランク付けされてユーザに提示され、さらにこの後、ユーザがこのように提示されたドキュメントの十分な数を確認するまで、次のグループが続き、以下同様である。図２はリモートサーバからドキュメントレコードおよびドキュメントを獲得するためにネットワーク接続を例示的に利用するものとして我々の発明を示すが、我々の発明はこのようには限定されない。図９Ａに関連して以下に詳細に説明するように、検索アプリケーションおよび我々の発明が共通のコンピュータ、すなわちローカルＰＣ上で実行され、そこにたとえばＣＤ－ＲＯＭまたは他の適切な媒体に記憶された付随するデータセットが配置されアクセス可能であれば、このようなネットワーク接続は必要ない。

【００２９】

図３は、図２に示されるコンピュータシステム３００のブロック図を示し、こ

のコンピュータシステム３００はは本発明の教示を取入れたものである。

【００３０】

示されるように、例えばクライアントパーソナルコンピュータであるこのシステムは、全て従来からバス３７０によって相互接続されている入力インターフェイス（ＩＮＰＵＴ　Ｉ／Ｆ）３３０と、プロセッサ３４０と、通信インターフェイス（ＣＯＭＭ　Ｉ／Ｆ）３５０と、メモリ３７５と、出力インターフェイス（ＯＵＴＰＵＴ　Ｉ／Ｆ）３６０とを含む。メモリ３７５は例えばランダムアクセスメモリ（ＲＡＭ）およびハードディスク記憶装置である種々の様式（これらは全て簡略化のために特に示してはいない）を一般的に含むが、オペレーティングシステム（Ｏ／Ｓ）３７８と、アプリケーションプログラム４００とを記憶する。我々の発明の教示を実装するソフトウェアは典型的にはアプリケーションプログラム４００に組込まれ、この実施例では特にウェブブラウザ（図４に示される）に組込まれる。このオペレーティングシステムは、ウィンドウズＮＴオペレーティングシステム（ワシントン州レッドモンド（Redmond, Washington）のMicrosoft Corporation（登録商標「ウィンドウズＮＴ」も所有する）から現在入手可能である。）など、どのような従来のオペレーティングシステムによって実装されてもよい。Ｏ／Ｓ　３７８の構成要素であるプロセスは発明とは無関係であるため、各部分については説明しない。しかしながら、ブラウザ、従って我々の発明のソフトウェアを、オペレーティングシステム自体の中に組込むこともできる。しかし、例示および簡略化の目的のために我々は、ブラウザがオペレーティングシステムから分離可能であり、アプリケーションプログラム４００内にあると仮定する。アプリケーションプログラム４００はＯ／Ｓ　３７８の制御下で実行される。ウェブブラウザを含む実行アプリケーションプログラムの各々に対して、１つまたは２つ以上の別個のタスクのインスタンスが、ユーザが特定した各コマンド、典型的にはメニューとか、ツールバー内のアイコンなど選択可能なコマンドが利用可能な場合にユーザが入力でバイス３９０を適切に操作することによって対話的に入力されるコマンドに応答して、ユーザによって呼出され、付随する情報がディスプレイ３８０に提示される。

【００３１】

図３に示されるように、入来する情報は、例えば２つの外部ソースから生ずる。すなわち、たとえばインターネットおよび／またはイントラネットなどのネットワーク化された他の設備（いずれも全体的に図２にネットワーク２１０として示される）から、ネットワーク接続２０５を介して通信インターフェイス３５０（図３に示される）へ達するネットワーク空供給される情報、または経路３１０を介して専用の入力ソースから入力インターフェイス３３０へ達するものである。専用の入力は、たとえばローカルであると、リモートであると、または他の入力ソースであるとにかかわらず、外部のデータセットなどさまざまなソースから生ずる。入力インターフェイス３３０は経路３１０に接続され、入力情報の、種々の専用のソースの各々をコンピュータシステム３００に物理的に接続してインターフェイスするために必要である対応の電気接続を提供する適当な回路構成を含む。アプリケーションプログラム４００は、オペレーティングシステムの制御下で、ネットワーク接続２０５を介してリモートウェブサーバなどの外部ソースと、または経路３１０を介して専用のソースなどと、コマンドおよびデータを交換し、プログラムの実行時に典型的にはユーザによってリクエストされる情報の送受信を行なう。

【００３２】

入力インターフェイス３３０は、リード３９５によって、キーボードおよびマウスなどのユーザ入力デバイス３９０をコンピュータシステム３００に電気的に接続し、インターフェイスする。従来のカラーモニタなどのディスプレイ３８０および従来のレーザプリンタなどのプリンタ３８５は、それぞれリード３６３および３６７によって出力インターフェイス３６０に接続される。出力インターフェイスはディスプレイおよびプリンタをコンピュータシステムに電気的に接続してインターフェイスさせるために不可欠な回路構成を提供する。実行中のアプリケーションからのハードコピー出力情報はプリンタ３８５によってユーザに与えられる。特に、ディスプレイ、プリンタおよび入力デバイス３９０（特定的にはマウスおよびキーボード）を適切に操作することにより、システム３００に配置されているユーザは、たとえば、インターネットを介して、そこからさらにアクセス可能なサーチエンジンを含む、膨大な数のリモートウェブサーバのうちのい

ずれかと画面を使って通信し、ローカルに表示および印刷するためにそこからドキュメントなどの情報を引出すことができる。

【００３３】

本発明の実現に必要なもの以外の、コンピュータシステム３００の特定のハードウェア構成要素およびメモリ３７５に記憶されたソフトウェアのそれぞれの局面は従来からのものであり周知であるため、これ以上詳細には説明しない。

【００３４】

図４は、図３に示されるコンピュータ３００内で実行されるアプリケーションプログラム４００の非常に高いレベルのブロック図を示す。これらのプログラムは本発明に関連する範囲では、図４に示されるようにウェブブラウザ４２０を含み、このウェブブラウザ４２０は、我々の本発明を実現するための検索プロセス６００（図６Ａおよび図６Ｂに関連して後に詳細に説明される）を含む。ウェブブラウザと、ＡＬＴＡ　ＶＩＳＴＡサーチエンジンなどのユーザが選択した統計的サーチエンジンとの間にインターネット接続が確立されているものと想定すると、ユーザは、図４に示される線４２２で表わされるように、プロセス６００にフルテキスト（「リテラル」）サーチ用のクエリを与える。このプロセスは、線４２６で表わされるように、ウェブブラウザを介してサーチエンジンにクエリを転送する。さらに、特に示してはいないが、プロセス６００はさらにクエリを内部で分析し、その対応の論理形式三つ組を生成し、これらは後にコンピュータ３００内にローカルに記憶される。クエリに応答して、サーチエンジンは線４３２で表わされるように統計的に検索されたドキュメントレコードの集合をプロセス６００に与える。これらのレコードの各々は、上述のとおり、のドキュメントをアクセスすることができるウェブアドレス、特定的にはＵＲＬと、さらに、そのドキュメントがあるリモートウェブサーバによって要求される、そのドキュメントを含むコンピュータファイルをインターネットを介してダウンロードするのに十分な、適切なコマンドを含む。プロセス６００がレコード全てを受信すると、このプロセスは、ウェブブラウザ４２０を介して、かつ線４３６によって表わされるように適切なコマンドを送信し、レコードによって特定された全てのドキュメント（すなわち出力ドキュメントの集合を形成するもの）にアクセスしてそれ

らをダウンロードしようとする。そしてこれらのドキュメントは、それらに対応するウェブサーバから順次アクセスされ、線４４２で表わされるように、ウェブブラウザ４２０に、特定的にはプロセス６００にダウンロードされる。これらのドキュメントがダウンロードされると、プロセス６００はこのようなドキュメントの各々を分析して、それに関する対応の論理形式三つ組を生成してローカルに記憶する。その後、各ドキュメントに関する論理形式三つ組に対して、クエリに関する論理形式三つ組を比較することにより、プロセス６００は少なくとも１つの一致する論理形式三つ組を含む各ドキュメントをスコア付けし、それらのスコアに基づいてこれらの特定のドキュメントをランク付け、最後に、線４４６によって表わされるように、ドキュメントのスコアの降順に、グループ毎にこれらの特定のドキュメントをユーザに提示するようにブラウザ４００に対して指示する。ブラウザ４００はディスプレイ３８０（図３参照）のスクリーン上に適切な選択ボタンを作成し、これによりユーザは、ユーザのマウスでそれを適切に「クリック」することによって選択を行ない、所望にしたがって、後続するドキュメントのグループの各々を表示することができる。

【００３５】

意味論上の情報を判断し、保存し、符号化する際の論理形式の有用性を十分に評価するために、この時点で、我々の発明を実現する処理の説明から離れ、関連する範囲で本発明に用いられる論理形式および論理形式三つ組を例示して説明し、それらが生成される態様を簡単に説明する。

【００３６】

大まかに言って、論理形式は、任意のサイズを有するテキストを表わす単語がラベル付けされた関係によってリンクされる、有向非循環グラフである。論理形式は、句内の重要な単語、この単語には上位語および／またはその類義語を含めることもあり得るが、これら単語間の意味論的な関係を描く。図５Ａから図５Ｄを参照して説明され例示されるように、論理形式は種々の多くの形態のうちいずれかをとることができ、たとえば論理形式三つ組のリストなどの、論理形式グラフまたは任意のそのサブグラフの形式をとることができる。たとえばこれら三つ組の各々は「語－関係子－語」という形式を持つ。この実施例では、本発明は論

理形式三つ組を生成して比較するが、本発明は、単語間の意味論的な関係を描くことができるものであれば、上述したような他のいかなる形式を容易に利用することができる。

【００３７】

論理形式三つ組およびそれらの構造は順により複雑になるような一連の文を例とすることによって最もよく理解できるため、まず図５Ａを参照する。この図は、例示的な入力文字列、特定的には「The octopus has three hearts.」という文に関する論理形式グラフ５１５と論理形式三つ組５２５とを示す。

【００３８】

一般に、入力文字列、たとえば入力文字列５１０に対する論理形式三つ組を生成するために、この文字列はまずパージングされてその構成要素の単語に分解される。その後、このような各単語に対して、予め格納された辞書にある、予め定義されたレコード（サーチエンジンによって採用されるドキュメントレコードと混乱してはならない。）を用いて、これらの構成要素の単語に対応のレコードが、予め規定された文法規則によって組み合わせられて大きな構造または構文になり、さらにそれらは予め定められた文法規則によって再度組み合わせられて、構文解析木などのさらに大きな構造を形成する。その後論理形式グラフが解析木から構築される。特定の規則が特定の構成要素の集合に適用可能であるか否かは、部分的には、単語レコードに、ある対応の属性およびそれらの値が存在するか否かによって支配される。そしてこの論理形式グラフは、一連の論理形式三つ組に変換される。例えば、我々の発明はおよそ１６５，０００個のヘッド単語のエントリを有する辞書を使用する。この辞書は入力文字列に関する解析木が構築できるように、入力文字列内の単語に固有な統語論上および意味論上の特性を規定する、前置詞、接続詞、動詞、名詞、オペレータおよび数量詞などの、さまざまなクラスの単語を含む。明らかに、論理形式（またはその問題に関しては、意味論上の関係を描くことができる、論理形式内の論理形式三つ組または論理形式グラフ）を、後に対応のドキュメントが検索されたときに計算するのではなく、対応のドキュメントが索引付けされている間に、予め計算してたとえばそのドキュメントに関するレコード内に格納して、後のアクセスに使用するようにできる。図

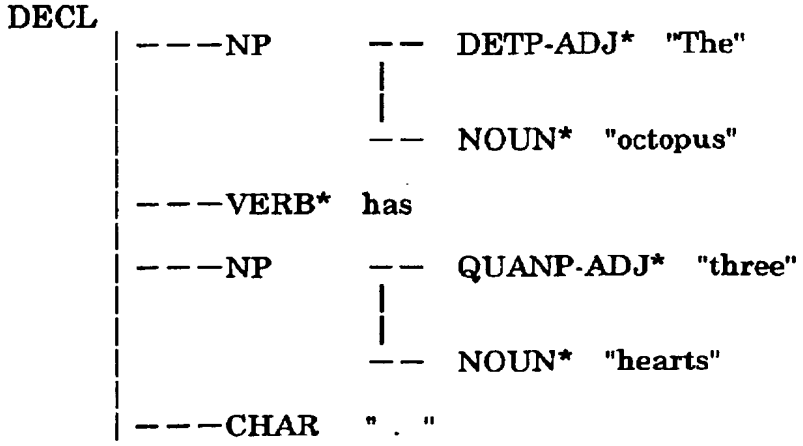
１０から図１３Ｂに関連して後に説明する我々の発明の別の実施例に見られるように、このように予め計算して格納しておくと、自然言語処理の量が劇的に低減するという効果があり、したがって、我々の発明に従って、検索されたドキュメントを扱うために必要な関連の実行時間が短くなる。

【００３９】

特に、図５Ａに示される文５１０などの入力文字列は、まず、その構成要素の単語の各々に関して辞書内にある予め規定されたレコードを用いて形態素分析され、それに関するいわゆる「語幹」（または「基体」）形式を生成する。語幹は、たとえば動詞の時制および単数－複数といった名詞の変形などの種々の単語の形を、パーサが使用できるような共通の形態素的形式に正規化するために用いられる。一旦語幹形式が作成されると、文法規則および構成要素の単語のレコード内にある属性を用いて、入力文字列がパーサによって構文解析され、それに関する構文解析木が生成される。この木は入力文字列の構造を示し、具体的には、たとえば入力文字列内の「The octopus」という名詞句のような各単語または句と、たとえば名詞句に対するＮＰのような、対応の文法的機能のカテゴリと、その中の、構文的に関連した各単語または句へのリンクとを表わす。例示的な文５１０については、関連の構文解析木は以下のとおりであろう。

【００４０】

【表１】



【００４１】

表１－－「The octopus has three hearts.」に関する構文解析木

木の上部左側にある開始ノードは、パージングされる入力文字列のタイプを定義する。文のタイプには、平叙文に関する「ＤＥＣＬ」（上の例）と、命令文に関する「ＩＭＰＲ」と、疑問文に関する「ＱＵＥＳ」とが含まれる。開始ノードの右下垂直に表示されるのは、第１レベルの構文である。この構文は、典型的には主動詞（この例では「has」という単語）である、星印によって示されたヘッドノードと、（この例では「The octopus」という名詞句である）前置修飾語句と、その後にくる（「three hearts」という名詞句である）修飾語句とを有する。木の葉の各々は辞書に含まれる単語または句読点を含む。ここでは、ラベルとして「ＮＰ」は名詞句を示し、「ＣＨＡＲ」は句読点を示す。

【００４２】

そして構文解析木は、異なった組の規則を用いてさらに処理され、入力文字列５１０に関するグラフ５１５などの論理形式グラフが生成される。論理形式グラフを作成するプロセスは、入力文字列の構文解析から下層の構造を抽出することを含み、論理形式グラフは、互いの間に意味関係と、その関係の機能的性質を有すると定義された複数の単語を含む。種々の意味関係の分類に使用される「深層」の格（Case）すなわち機能的役割は、以下を含む。

【００４３】

Ｄｓｕｂ――深層主語

Ｄｉｎｄ――深層間接目的語

Ｄｏｂｊ――深層目的語

Ｄｎｏｍ――深層述語主格

Ｄｃｍｐ――深層目的格補語

表２

入力文字列内の意味関係全てを特定するために、その文字列に関する構文解析木の各ノードが検査される。上記の関係に加えて、たとえば下記の他の意味的役割が使用される。

【００４４】

ＰＲＥＤ　――述部

ＰＴＣＬ　――２部構成の動詞における不変化詞

Ｏｐｓ　　－－機能語、たとえば数字

Ｎａｄｊ　－－名詞を修飾する形容詞

Ｄａｄｊ　－－叙述形容詞

ＰＲＯＰＳ－－節である、他には特定されない修飾部

ＭＯＤＳ　－－節でない、他には特定されない修飾部

表３

同様に追加の意味的ラベルが下記のように規定される。

【００４５】

ＴｍｅＡｔ－－時刻

ＬｏｃＡｔ－－場所

表４

いずれにせよ、入力文字列５１０に関するこのような分析の結果は論理形式グラフ５１５である。入力文字列中の単語で互いの間に（たとえば「Octopus」と「Have」との間の）意味関係が認められる単語は、互いにリンク付けられ、それらの間の関係はリンク付け属性（たとえばＤｓｕｂ）として特定されて示されている。このグラフは、入力文字列５１０に関するグラフ５１５に代表されるように、各入力文字列に関する主題および修飾語句の構造を捕らえている。とりわけ、論理形式分析は、前置詞および冠詞などの機能的単語を、グラフ内に示された特徴または構造上の関係にマッピングする。論理形式分析はさらに前方照応を解決、すなわち、たとえば代名詞と、同一指示名詞句との間の正しい先行関係を規定し、さらに省略に関する適切な機能的関係を検出して示す。論理形式分析時には、曖昧さおよび／または他の言語的特異性に対処するためにさらに処理が施されることもあり得る。そして、対応の論理形式三つ組が従来の態様で論理形式グラフから読出され、組として記憶される。各三つ組は、グラフに示されるように、互いの間の意味関係によってリンク付けられた２つのノード単語を含む。例としての入力文字列５１０に関しては、論理形式三つ組５２５が処理グラフ５１５から結果として得られる。ここでは、論理形式三つ組５２５は、入力文字列５１０に固有の意味論的な情報を全体として伝える３つの別個の三つ組を含む。

【００４６】

同様に、図５Ｂから図５Ｄに示されるように、入力文字列５３０、５５０および５７０、すなわち例示の文「The octopus has three hearts and two lungs.」、「The octopus has three hearts and it can swim.」、および「I like shark fin soup bowls.」に対しては、論理形式グラフ５３５、５５５および５７５ならびに論理形式三つ組５４０、５６０および５８０がそれぞれ結果として得られる。

【００４７】

論理形式三つ組が論理形式グラフから生成される従来の「グラフウォーク」を含む従来の態様のものとは別に、論理形式三つ組の全てを正しく得るために必要な追加の自然言語処理が必要な、３つの論理形式構造がある。例示の文「The octopus has three hearts and two lungs.」、すなわち入力文字列５３０におけるような等位語の場合、単語と、その意味関係と、等位された構成要素の値の各々とに対する論理形式三つ組が生成される。「特殊な」グラフウォークによると、図５４０には２つの論理形式三つ組「have-Dobj-heart」および「have-Dobj-lung」があることがわかる。従来のグラフウォークのみを用いると、１つの論理形式三つ組「have-Dobj-and」しか得られなかったであろう。同様に、指示語（Refs）を有する構成要素の場合、例示の文「The octopus has three hearts and it can swim.」、すなわち入力文字列５５０の場合の様に、従来のグラフウォークによって生成された三つ組の他に、単語と、その意味関係と、Refs属性の値の各々とに対する論理形式三つ組を生成する。この特殊なグラフウォークによると、従来の論理形式三つ組「swim-Dsub-it」の他に論理形式三つ組「swim-Dsub-octopus」が三つ組５６０に見出される。最後に、例示の文「I like shark fin soup bowls.」、すなわち入力文字列５７０におけるように、名詞の修飾語で構成される場合、名詞の複合語の可能な内部構造を表わすためにさらなる論理形式三つ組が生成される。従来のグラフウォークでは、可能な内部構造[[shark] [fin] [soup] bowl]を反映する論理形式三つ組「bowl-Mods-shark」、「bowl-Mods-fin」および「bowl-Mods-soup」が生成される。特殊なグラフウォークの場合、下記の可能な内部構造[[shark fin] [soup] bowl]、[[shark] [fin soup] bowl]および[[shark [fin] soup] bowl]を表わすためにそれぞれ追加の論理形式三つ組

「fin-Mods-shark」、「soup-Mods-fin」、および「soup-Mods-shark」が生成される。

【００４８】

形態素的、統語論的および論理形式処理の特定的な詳細は本発明とは関係ないため、さらに詳細な説明は省略することとする。しかしながら、これに関するさらなる詳細は、１９９６年６月２８日に出願され、連続番号第０８／６７４，６１０号が付与された「構文解析木から意味論的論理形式を計算するための方法およびシステム（"Method and System for Computing Semantic Logical Forms from Syntax Trees"）」と題された同時係属中の米国特許出願と、特に１９９７年３月７日に出願された、連続番号＿＿＿＿＿が付与された「テキストの意味表現を利用する情報検索（"Information Retrieval Utilizing Semantic Representation of Text"）」とを参照されたい。これらはいずれも本願の譲受人に譲渡されており、引用によってここに援用される。

【００４９】

この論理形式の概要およびそれらの構造を念頭に、我々の本発明を実現する処理の議論に戻ることとする。

【００５０】

図２、図３、図４に示される我々の発明の特定の実施例に使用されている、我々の発明の検索プロセス６００のフローチャートは、図６Ａおよび図６Ｂにまとめて示され、これらの図面の正確な配列が図６に示される。破線で描かれたブロック２２５に示される動作以外の、これらの図面に示される残りの動作は、たとえばクライアントＰＣ３００（図２および図３参照）であるコンピュータシステムによって行なわれ、具体的にはウェブブラウザ４２０内で行なわれる。理解を容易にするために、以下の説明を読みながら図２、図３、図６Ａおよび図６Ｂを同時に参照すべきである。

【００５１】

プロセス６００に入ると、実行処理はまずブロック６０５に進む。このブロックは、実行されると、フルテキスト（リテラル）のクエリをウェブブラウザ４２０に入力することをユーザーに促す。クエリは単一の質問（たとえば「Are ther

e any air-conditioned hotels in Bali?」）でも、単一の文（たとえば「Give me contact information for all fireworks held in Seattle during the month of July.」）でも、文の一部（たとえば「Clothes in Ecuador」）の形式であってもよい。この質問が得られると、実行処理は経路６０７を介してブロック６１０に、および経路６４３を介して経路６４５に分岐する。ブロック６４５が実行されると、それはＮＬＰルーチン７００を呼出し、クエリを分析し、その対応の論理形式三つ組の集合を構築してローカルに記憶する。ブロック６１０が実行されると、それは、破線６１５で表わされるように、インターネット接続によって、フルテキストのクエリを、ウェブブラウザ６２０から、サーバ２２０に置かれたエンジン２２５などのリモートサーチエンジンに送信する。この時点で、ブロック６２５がサーチエンジンによって実行され、クエリに応答してドキュメントレコードの集合を取出す。この集合が形成されると、破線６３０で表わされるように、その集合はリモートサーバによってコンピュータシステム３００に再送信され、特に、そこで実行されているウェブブラウザ４２０に戻される。その後、ブロック６３５が実行され、レコードの集合を受信し、後に、各レコードに対して、そのレコードからＵＲＬを抽出し、そのＵＲＬでウェブサイトにアクセスし、さらにはそのレコードに対応するドキュメントを含む関連ファイルをそこからダウンロードする。全てのドキュメントがダウンロードされると、ブロック６４０が実行される。このような各ドキュメントに対して、このブロックはまず、そのドキュメントに関連するＨＴＭＬタグ内にある全てのテキストを含む、全てのテキストをそのドキュメントから抽出する。その後、一度に一つの文に対して行なわれる自然言語処理を容易にするために、各ドキュメントに関するテキストが、従来の一文切出し処理によって各文章（または質問）がファイル内の別々の行を占めるようなテキストファイルに切出される。その後、ブロック６４０は、そのドキュメントのテキストの各行に対して、ＮＬＰルーチン７００（これは図７に関連して後に詳細に説明される）を繰返して呼出し、これらのドキュメントの各々を分析し、そのドキュメントのテキストの各行に関する対応の論理形式三つ組の集合を構築してローカルに記憶する。ブロック６４５における動作は、ブロック６１０、６３５および６４０におけるものと基本的に並行して行なわれ

るものとして説明したが、前者のブロックの動作は、実際の実装での条件に基づいて、ブロック６１０、６３５および６４０の動作と順次に、それらの前または後のいずれかに行なってもよい。これに代えて、図１０から図１３Ｂに関連して後に説明する我々の発明の別の実施例の場合のように、各ドキュメントに関する論理形式三つ組を予め計算して記憶しておき、後のドキュメントの検索時に使用してもよく、この場合にはこれらの三つ組はドキュメントの検索時には計算されずに単にアクセスされるだけである。この場合、これら三つ組は、何らかの態様で、その格納されたドキュメントのプロパティ（属性）として、または、たとえばそのドキュメントに関するレコードまたはそのドキュメントを含むデータセットのいずれかに別のエントリとして、記憶されるであろう。

【００５２】

いずれにせよ、図６Ａおよび図６Ｂに示されるプロセス６００に戻り、論理形式三つ組の集合がクエリおよび出力ドキュメントの集合内の検索されたドキュメントの各々の双方に対して構築され完全に記憶されると、ブロック６５０が実行される。このブロックは、クエリの論理形式三つ組の各々と、検索されたドキュメントの各々に関する論理形式三つ組の各々とを比較して、クエリのいずれかの三つ組と、ドキュメントのいずれかの三つ組のいずれかとの一致を突き止める。例としての一致の形式は、これらの三つ組の間での、ノード単語と、関係子タイプとの双方の点においてこれら２つの三つ組間で同一の一致が見られることと定義される。特に、例示の１対の論理形式三つ組、すなわち「語１ａ－関係子１－語２ｉおよび語１ｄ－関係子２－語２ｂ」の場合、ノード単語の語１ａおよび語１ｂが同一であり、ノード単語の語２ａおよび語２ｂが同一であり、関係子１および関係子２が同じである場合にのみ一致が起こる。１つの三つ組の３つの要素の全てが別の三つ組の対応する要素と同一に一致しないならば、これらの２つの三つ組は一致しない。ブロック６５０が完了すると、ブロック６５５が実行され、一致する三つ組が得られない、すなわちクエリの三つ組と一致する三つ組がない、検索された全てのドキュメントが破棄される。その後、ブロック６６０が実行される。ブロック６６０により、一致する三つ組の関係子のタイプ、およびこれらのドキュメントの各々に関して存在するそれらの重みに基づいて、残りのド

キュメント全てにスコアが割当てられる。特に、論理形式三つ組に生じ得る異なったタイプの関係子の各々に、図８Ａの表８００に示されるもののような対応の重みが割当てられる。たとえば、図示されるように、例示の関係子Dobj、Dsub、OpsおよびNadjにはそれぞれ、１００、７５、１０および１０という、予め定められた静的な数の重みを割当てることができる。重みはクエリとドキュメントとの正しい意味上の一致を示す上で、その関係子に帰すると考えられる相対的な重要性を反映する。これらの重みの実際の数値は一般に、経験に基づいて定義される。後に図８Ｂに関連して詳細に説明するように、残りのドキュメントの各々に関して、そのスコアは予め定義された関数であり、ここでは例として、その一意に一致する三つ組（二重に一致する三つ組は全て無視する。）の重みの数値の和である。こうしてドキュメントが一旦重み付けされると、ブロック６６５が実行されて、スコアの降順に、ドキュメントをランク付ける。最後に、ブロック６７０が実行されて、典型的には、最も高いスコアを示す予め規定された小さなグループのドキュメント、典型的には５個から１０個のドキュメントをランク順に表示する。その後、ユーザーは、たとえばウェブブラウザ４２０によって表示された対応のボタン上でユーザーのマウスを適当に「クリック」することにより、コンピュータシステム（クライアントＰＣ）３００に、ランク付けされたドキュメントの次のグループを表示させ、以下同様に、ユーザーがランク付けされたドキュメント全てを順次十分に検討するまでこれを続け、そこでプロセス６００を完了する。

【００５３】

図７は、ＮＬＰルーチン７００のフローチャートを示す。このルーチンは、１行の入力テキストを与えられ、そのテキストがクエリ、ドキュメント内の文、またはテキストの一部のいずれの場合にも、それに関する対応の論理形式三つ組を構築する。

【００５４】

特に、ルーチン７００に入ると同時に、ブロック７１０はまず入力テキストの行を処理し、図５Ａに示される例示のグラフ５１５などの論理形式グラフを生成する。この処理は、構文解析木を生成する形態素的および統語論上の処理を含み

、この構文解析木から後に論理形式グラフが算出される。その後、図７に示されるように、ブロック７２０が実行されて、グラフから対応の論理形式三つ組の集合を抽出する（読出す）。これが行なわれると、ブロック７３０が実行されて、このような論理形式三つ組の各々を、別個で区別された、フォーマット化されたテキスト文字列として生成する。最後に、ブロック７４０が実行されて、データセット（すなわちデータベース）に入力テキストの行を記憶し、フォーマット化された一連のテキスト文字列として、その行に関する論理形式三つ組の集合を記憶する。この集合が完全に記憶されると、実行処理はブロック７００を出る。これに代えて、論理形式三つ組の代わりに、たとえば論理形式グラフのような、異なった表現が論理形式に関連付けられて我々の発明に関し用いられる場合には、その特定の形式をフォーマット化された文字列として生成するようにブロック７２０および７３０を容易に変更でき、データセットへの論理形式三つ組の代わりにその形式を記憶するようにブロック７４０を変更できる。

【００５５】

論理形式三つ組の一致を比較しそれに重み付けをし、さらには対応のドキュメントをランク付ける、我々の発明の態様を十分に認識するために、図８Ｂを参照する。この図は、我々の発明の教示に従う論理形式三つ組の比較、ドキュメントの記憶、ランク付けおよび選択処理を図で示し、これらは、例示のクエリおよび検索された３つのドキュメントの例示の組に関する、図６Ａおよび図６Ｂに全て示されるブロック６５０、６６０、６６５および６７０において行なわれる。例示の目的で、ユーザーがフルテキストのクエリ８１０を我々の発明の検索システムに与えたと仮定し、このクエリが「How many hearts does an octopus have?」というものであるとする。さらに、このクエリに応答して、統計的サーチエンジンによって最終的に３つのドキュメント８２０が検索されたものとする。これらのドキュメントのうち、第１のドキュメント（ドキュメント１と記す。）はartichoke heartsおよびoctopusを含むレシピである。第２のドキュメント（ドキュメント２と記す。）はoctopi（タコ一般）に関する論文である。第３のドキュメント（ドキュメント３と記す。）はdeer（鹿）に関する論文である。これらの３つのドキュメントおよびクエリはそれらの構成要素の論理形式三つ組に変換さ

れ、それらに関する処理は包括的に「ＮＬＰ」（自然言語処理）によって表わされる。結果として得られる、クエリとドキュメント１、ドキュメント２およびドキュメント３とに関する論理形式三つ組は、それぞれブロック８３０、８４０、８５０および８６０に与えられる。

【００５６】

これらの三つ組が一旦こうして定義されると、破線８４５、８５５および８６５で表わされるように、クエリに関する論理形式三つ組がそれぞれドキュメント１、ドキュメント２およびドキュメント３に関する論理形式三つ組と順次比較され、いずれかのドキュメントが、クエリのいずれかの論理形式三つ組と一致する三つ組を含むか否かが確認される。ドキュメント１の場合のように、このような一致する三つ組を含まないドキュメントは破棄され、さらに考慮されない。一方、ドキュメント２およびドキュメント３は一致する三つ組を含む。特に、ドキュメント２はこのような三つ組を３つ含む。すなわち、これらはたとえば１つの文に関連する「HAVE-Dsub-OCTOPUS」および「HAVE-Dsub-HEART」と、たとえば別の文に関連する「HAVE-Dsub-OCTOPUS」である（これらの文は特定的には示さない）。これらの三つ組のうち２つは同一であり、すなわちそれは「HAVE-Dsub-OCTOPUS」である。ドキュメントに関するスコアは例えば、そのドキュメント内の、一意に一致する三つ組全ての重みの数値の和である。全てのドキュメントに関し、二重に一致する三つ組は全て無視される。三つ組に生じ得る異なったタイプの関係子の相対的な重みの例示的なランク付けは、最も大きな重みから小さな重みの順に、最初が動詞－目的語の組合せ（Dobj）、動詞－主語の組合せ（Dsub）、前置詞および機能語（たとえば（Ops）、および最後には修飾語句（たとえばNadj）である。このような重み付け方式を、図８Ａに示される例示の三つ組重み付け表８００に示す。この図を簡単にするために、表８００は論理形式三つ組に生じ得る種々の関係子全てを含むのではなく、図８Ｂに示される三つ組に関連するものだけを示す。この測定基準により、各ドキュメントのうちそのスコアに寄与する特定の三つ組にチェック（「レ」）マークを付与してある。もちろん、ドキュメントにスコア付けするための予め規定される測定基準として我々が選択したもの以外を用いてもよく、たとえば、ドキュメントの選択性（区別）を高めるた

めに重みを加算するのではなく乗算すること、または重みを別の態様って加算すること、たとえば同じタイプの複数の一致を含めることおよび／または上述以外の他の三つ組の重みを除くことなどであってもよい。さらに、任意のドキュメントに関して、スコアはある態様で下記のものを考慮に入れるであろう。すなわち、そのドキュメントにおける三つ組自体のノード単語、そのドキュメント内のこれらのノード単語の頻度または意味的な内容、そのドキュメント内の特定のノードの単語の頻度または意味的な内容、またはそのドキュメント内の特定の論理形式（またはそれのパラフレーズ）および／または特定の論理形式三つ組全体としての頻度、ならびにそのドキュメントの長さである。

【００５７】

したがって、我々が選択したスコア付けの測定基準および図８Ａの表８００に明記される重みを考慮すると、ドキュメント２のスコアは１７５であり、これはドキュメント内の、ブロック８５０に示される第１の文に関連した最初の２つの三つ組の重み、すなわち１００および７５を組合せることにより形成される。このドキュメント内の、その第２の文に関連し、このブロックに記載されている第３の三つ組であって、ドキュメント内に存在する他の三つ組の１つと既に一致しているものは無視される。同様に、ドキュメント３に関するスコアは１００であり、これは、ブロック８６０に記載されているようにこの特定のドキュメント内で唯一の一致する三つ組に関する重み、すなわちここでは１００により形成される。スコアに基づいて、ドキュメント２がドキュメント３より前にランク付けされ、これらのドキュメントはその順番でユーザーに提示される。ここでは起こらなかったが、いずれか２つのドキュメントが同じスコアを有する場合、これらのドキュメントは従来の統計的サーチエンジンによって提供されるのと同じ順番でランク付けされ、その順番でユーザーに提示される。

【００５８】

明らかに、当業者であれば、我々の本発明を実現するために使用される処理のさまざまな部分が、単一のコンピュータ内に存在しても、全体として情報検索システムを形成する種々のコンピュータに分散いてもよいことが容易に理解されるであろう。この点に関し、図９Ａから図９Ｃはそれぞれ、我々の本発明の教示を

採用した情報検索システムの異なった３つの実施例を示す。

【００５９】

このような代替的な実施例の１つが図９Ａに示され、ここでは全ての処理がＰＣなどの単一のローカルコンピュータ９１０によって行なわれる。この場合、コンピュータ９１０はサーチエンジンをホストし、そのエンジンによって、入力ドキュメントを索引付け、ユーザーによって与えられたフルテキストのクエリに応答して（ＣＤ－ＲＯＭまたは他の記憶媒体などによってそこにローカルに置かれるか、またはそのコンピュータにアクセス可能である）データセットをサーチし、出力ドキュメント集合を形成する、検索されたドキュメントの集合を最終的に生成する。このコンピュータはさらに我々の発明の処理をホストし、クエリおよびこのような各ドキュメントの両方を分析して、対応の論理形式三つ組の集合を生成し、その後三つ組の集合を比較し、上述の態様でドキュメントをスコア付けてランク付け、最後に、たとえばそこに配置されている、またはそこにアクセス可能なローカルユーザーに結果を提示する。

【００６０】

別の代替的な実施例が図９Ｂに示され、この図９Ｂは図２に示される特定の状況を包含するものであって、ここでは検索システムはリモートサーバにネットワーク接続されたクライアントＰＣで形成される。ここでは、クライアントＰＣ９２０はネットワーク接続９２５によってリモートコンピュータ（サーバ）９３０に接続される。クライアントＰＣ９２０にいるユーザーはフルテキストのクエリを入力し、ＰＣはこれをネットワーク接続を介してリモートサーバに送信する。クライアントＰＣはさらにクエリを分析して、その対応の論理形式三つ組の集合を生成する。サーバは、たとえば従来の統計的サーチエンジンをホストし、その結果このクエリに応答して、統計的検索を行ない、ドキュメントレコードの集合を生成する。そしてサーバはレコードの集合を戻し、最終的に、クライアントの命令によって、またはサーチエンジンまたは関連ソフトウェアの能力に基づいて自律的に、出力ドキュメントの集合内にある各ドキュメントをクライアントＰＣに戻す。そしてクライアントＰＣは出力ドキュメントの集合内の、受信した対応のドキュメントの各々を分析し、それに関する論理形式三つ組の集合を生成する

。クライアントＰＣは次いで、三つ組の集合を適切に比較し、上述の態様でドキュメントを選択し、スコア付けし、ランク付けし、最後に結果をローカルユーザーに提示して、その処理を完了する。

【００６１】

さらなる実施例を図９Ｃに示す。この実施例は図９Ｂと同じ物理的ハードウェアおよびネットワーク接続を用いるが、クライエントＰＣ９２０がローカルユーザからのフルテキストクエリの依頼を受入れ、そのクエリの依頼をネットワーク接続９２５を介してリモートコンピュータ（サーバ）９３０へ転送する。このサーバは単に従来のサーチエンジンをホストするのではなく、本発明に従う自然言語処理を提供する。この場合、クライエントＰＣではなくサーバがクエリを適切に分析してそのための論理形式三つ組の対応の集合を生じるであろう。サーバはまた必要であれば出力ドキュメント集合内の検索された各ドキュメントをダウンロードし、次にこのような各ドキュメントを分析してそのための論理形式三つ組の対応の集合を生成するであろう。その後、サーバはクエリのための三つ組の集合とドキュメントとを適切に比較し、前述のようにドキュメントを選択し、それにスコアを付け、ランクを付けるであろう。一旦このランク付けが行なわれると、サーバ９３０は残りの検索ドキュメントをランク順にネットワーク接続９２５を介してクライエントＰＣ９２０に送信し、そこで表示させるであろう。サーバはこれらのドキュメントを前述のようにユーザの指示に従ってグループごとに送信するか、それらをグループごとに選択してクライエントＰＣで表示させるために全てのドキュメントを順次送信することができる。

【００６２】

さらに、リモートコンピュータ（サーバ）９３０は、前述の従来の検索処理、自然言語処理および関連の処理の全てを与える１台のみのコンピュータによって実現される必要はなく、図９Ｄに示す分散処理方式であってもよい。その場合、このサーバが請け負う処理は分散処理方式における個別のサーバ間に分散される。ここで、サーバ９３０は、メッセージを接続９５０を介して（サーバ１、サーバ２、…、サーバｎを含む）一連のサーバ９６０に分散するフロントエンドプロセッサ９４０からなる。これらのサーバの各々が本発明のプロセスの特定の部分

を実施する。この点で、サーバ１は後の検索のために入力ドキュメントの索引付けを行ない大容量データ記憶装置上のデータセットに格納するために用いることができる。サーバ２は、フロントエンドプロセッサ９４０によって送られるユーザから与えられるクエリに応答して大容量データ記憶装置から一組のドキュメントレコードを引出すための従来の統計的エンジンのようなサーチエンジンを実現できる。これらのレコードは、サーバ２からフロントエンドプロセッサ９４０を介してたとえばサーバｎに送られ、対応のウェブサイトまたはデータベースからの対応の各ドキュメントを出力ドキュメント集合中にダウンロードするというような後処理が行なわれるであろう。フロントエンドプロセッサ９４０はまたクエリをサーバｎに送るであろう。サーバｎは次にそのクエリおよび各ドキュメントを適切に分析して論理形式三つ組の対応の集合を生じ、次に三つ組の集合を適切に比較し、前述のようにドキュメントを選択し、それにスコアをつけ、ランクをつけ、その後ランク付けされたドキュメントをフロントエンドプロセッサ９４０を介してクライエントＰＣ９２０に戻して、そこでランク付けの表示がされるようにする。もちろん、本発明の処理において用いられるさまざまな動作は、実行時に生ずる条件および／または他により生ずる条件次第で、静的であれ動的であれ、他の多くの方法のうちの任意の方法によってサーバ９６０中に分散されてもよい。さらに、サーバ９３０は、たとえば従来のサーチエンジンのためのデータベースと自然言語処理のために用いられる辞書との両方が記憶され、サーバ内の全てのプロセッサからアクセス可能な共用直接アクセス記憶装置（ＤＡＳＤ）である、たとえば周知のシスプレックス構成（または他の同様の分散マルチプロセス環境）によって実現することもできる。

【００６３】

本発明を、検索された各ドキュメントレコードに応答してドキュメントをダウンロードし、次にそのレコードをたとえばクライエントＰＣによってローカルに分析してその対応の論理形式三つ組を生じるものとして説明したが、これに替えてこれらの三つ組はドキュメントに対しサーチエンジンが索引付けをしている間に生成されてもよい。この点で、サーチエンジンがたとえばウェブクローラを用いて、索引付けを行なうための新しい各ドキュメントを見つけたときに、エンジ

ンがそのドキュメントのための完全なファイルをダウンロードし、それからその直後またはさらに後にバッチ処理でそのドキュメントを分析し、その論理形式三つ組を生成することによってそのドキュメントを前処理することができる。前処理の終了時にサーチエンジンは次にこれらの三つ組をそのドキュメントのための索引付けされたレコードの一部としてそのデータベースに記憶するであろう。後に、そのドキュメントレコードがたとえばサーチクエリに応答して検索されるたびに、そのための三つ組がドキュメントレコードの一部として比較などの目的のためにクライエントＰＣに戻される。サーチエンジン内でのドキュメントの前処理によって、クライエントＰＣでのかなりの量の処理時間が節約されるという効果があり、それによってクライエントのスループットを増大させることができる。

【００６４】

さらに、本発明をインターネットベースのサーチエンジンでの具体的な使用を例として説明したが、本発明は、（ａ）インターネットベースであろうとなかろうと、専用のネットワーク設備等によってアクセス可能な任意のネットワークアクセス可能なサーチエンジン、（ｂ）それ自身が所持する予め記録されたデータセットに対して動作するローカルなサーチエンジン、たとえば、百科事典、年鑑または他の独立型スタンドアローンデータセットに代表されるＣＤ－ＲＯＭベースのデータ検索アプリケーション、および／または（ｃ）その任意の組合せでの使用に等しく適用可能である。

【００６５】

上記を念頭において、図１０Ａおよび図１０Ｂは、本発明のさらに他の実施例を集合的に示し、この実施例ではドキュメントの前処理によって論理形式三つ組を発生し、結果として生じる三つ組、ドキュメントレコードおよびドキュメント自体を独立型スタンドアローンデータセットとして既存の記憶媒体、たとえば、１つ以上のＣＤ－ＲＯＭまたは（着脱可能なハードディスク、テープ、もしくは、光磁気または大容量磁気または電子記憶装置に代表される）他の可搬の大容量媒体にまとめて保存することによりエンドユーザへの頒布が容易になる。これらの図面の正しい配置は図１０に示すとおりである。検索アプリケーション自体と

それに付随するサーチされるべきデータセットとを共通の媒体にまとめて入れることによって、スタンドアローンのデータ検索アプリケーションが得られ、それによって、ドキュメントを検索するためにリモートサーバにネットワーク接続することが必要でなくなる。

【００６６】

図示したように、この実施例はドキュメント索引付け部分１００５$_1$、複製部分１００５$_2$およびユーザ部分１００５$_3$の本質的に３つの部分からなる。部分１００５$_1$はドキュメントを集め索引付けしてデータセット、すなわち図示するデータセット１０３０を作成し、データセット１０３０は、独立型ドキュメント検索アプリケーション、たとえば、百科事典、年鑑、（判例集のような）専用ライブラリ、定期刊行物のコレクション等のためのドキュメントリポジトリを形成する。大記憶容量を有するＣＤ－ＲＯＭおよび他の形態の媒体を複製するためのコストは急速に低下しつつあり、この実施例は大量のドキュメントをそれを正確にサーチする性能とともに広いユーザコミュニティに費用効率よく頒布するために特に魅力的である。

【００６７】

いずれにせよ、索引付けされデータセットを形成するために入力されるドキュメントは任意数の多様なソースから集められ、コンピュータ１０１０に順次与えられる。このコンピュータはメモリ１０１５内に記憶されている適切なソフトウェアによってドキュメント索引付けエンジンを実現し、ドキュメント索引付けエンジンは、このような各ドキュメントのためのレコードをデータセット１０３０内に作成し、そのドキュメントのためのレコードに情報を保存し、またドキュメント自体のコピーを含む適切なエントリをデータセット内に作成し保存する。エンジン１０１５は三つ組発生プロセス１１００を実行する。図１１に関連して以下に詳細に説明するこのプロセスは索引付けされる各ドキュメントごとに別個に実行される。本質的には、このプロセスは、図６Ａおよび図６Ｂに示すブロック６４０に対して前述したのと本質的に同じように、ドキュメント内のテキスト句を分析し、そうすることによってそのドキュメントに対する論理形式三つ組の対応の集合を構成してデータセット１０３０内に記憶する。図１０Ａおよび図１０

Ｂに示す索引付けエンジン１０１０によって実行されてドキュメントに索引を付ける他の全てのプロセス、たとえばそのための適切なレコードを発生するプロセスはいずれも本発明には無関係であるので、それらについては詳細に述べない。三つ組の集合がプロセス１１００によって一旦発生されると、エンジン１０１５がこの集合をドキュメント自体のコピーとそれに対して作られたドキュメントレコードとともにデータセット１０３０へと記憶する、と述べるだけで十分である。したがって、全索引付け動作が終わると、データセット１０３０は索引付けされた全ドキュメントの完全なコピーとそのためのとをその中に記憶しているだけでなく、そのドキュメントのための論理形式三つ組の集合をも記憶している。

【００６８】

一旦所望の全ドキュメントが適切に索引付けされると、データセット１０３０は、これは「マスタデータセット」と見ることができるが、次に複製部分１００$5_2$によって複製される。部分１００$5_2$内では、従来の媒体複製システム１０４０が線１０３５によって供給されるマスタデータセットの内容のコピーを、線１０４３によって供給される検索プロセスおよびユーザインストールプログラムを含む検索ソフトウェアのための適切なファイルのコピーとともに、１つ以上のＣＤ－ＲＯＭのような共通の記憶媒体に繰返し書込んで、スタンドアローンのドキュメント検索アプリケーションをまとめて形成する。システム１０４０によって、個々の複製1050_1、1050_2、…、１０５０ｎを有する一連の媒体複製１０５０が生成される。全複製は同一であり、複製1050_1に関して具体的に示してあるように、線１０４３によって供給されるドキュメント検索アプリケーションファイルのコピーと線１０３５によって供給されるデータセット１０３０のコピーとを含む。データセットのサイズおよび構成次第では、各複製が１つ以上の別個の媒体、たとえば別個のＣＤ－ＲＯＭにまたがってもよい。後に、複製は典型的にはライセンスの取得によって破線１０５５で示すようにユーザコミュニティ中に流通される。

【００６９】

一旦ユーザ、たとえばユーザ$_j$がユーザ部分1005_3に示すように（ＣＤ－ＲＯＭ１０６０とも示す）$CD-ROM_j$のような複製を入手すると、ユーザは、

（同一の構成でないとしても実質的に図３に示すクライエントＰＣ３００のような構成を有するＰＣのような）コンピュータシステム１０７０によって、本発明を含むドキュメント検索アプリケーションをＣＤ－ＲＯＭjに記憶されているデータセットに対して実行してそこから所望のドキュメントを引出すことができる。特に、ユーザはＣＤ－ＲＯＭjを入手した後、ＣＤ－ＲＯＭをＰＣ１０７０に挿入し、ＣＤ－ＲＯＭに記憶されているインストールプログラムの実行を始め、それによって、ドキュメント検索アプリケーションファイルのコピーを作り、それをＰＣのメモリ１０７５、通常はハードディスク内の予め規定されたディレクトリへとインストールし、それによって、ＰＣ上にドキュメント検索アプリケーション１０８５を作成する。このアプリケーションはサーチエンジン１０９０および検索プロセス１２００を含む。一旦インストールが完了し、アプリケーション１０８５が呼出されると、ユーザは適切なフルテキストのクエリをアプリケーションに与えることによって、ＣＤ－ＲＯＭjのデータセットをサーチすることができる。クエリに応答して、サーチエンジンはそれらのドキュメントのためのとこのような各ドキュメントのための記憶されている論理形式三つ組とを含むドキュメントの集合をデータセットから引出す。クエリは検索プロセス１２００にも与えられる。このプロセスは図６Ａおよび図６Ｂに関連して前述した検索プロセス６００に非常に類似しており、クエリを分析し、そのため論理形式三つ組を構成するものである。その後、図１０Ａおよび図１０Ｂに示すプロセス１２００がその集合内の検索されたドキュメントの各々のための論理形式三つ組、特にそのためのレコードをクエリのための三つ組と比較する。それらの間で発生する三つ組の一致とそれらの重みとに基づき、プロセス１２００は詳細に前述した態様で少なくとも１つの一致する三つ組を示すドキュメントの各々をスコア付けし、これらのドキュメントを降順のスコアでランク付けし、最後に、最も高いランク付けを有する典型的に５－２０またはそれよりも少ない小グループのドキュメントレコードをユーザに視覚的に提示する。ユーザはこれらのを検討し、関連のあるように思われる任意のドキュメントのコピー全体を検索し表示するようドキュメント検索アプリケーションに指示することができる。一旦ユーザが最初のグループの検索ドキュメントに対する最初のグループのドキュメントレコードを検討

すると、ユーザは次に高いランク付けを有する次のグループのドキュメントレコードを要求することができ、以下同様に、検索された全ドキュメントレコードを検討し終わるまでこれを行なうことができる。アプリケーション１０８５は、初期状態では、ランク付けされたドキュメントレコードをクエリに応答して戻すが、これに替えてこのアプリケーションがドキュメント自体のランク付けされたコピーをクエリに応答して戻してもよい。

【００７０】

図１１は、図１０Ａおよび図１０Ｂに示すドキュメント索引付けエンジン１０１５によって行なわれる三つ組発生プロセス１１００を示す。前述のように、プロセス１１００は索引付けされるべきドキュメントの前処理を、そのドキュメントにおけるテキストフレーズを分析し、そうすることによってそのドキュメントのための論理形式三つ組の対応の集合を構成してデータセット１０３０内に記憶することによって行なう。特に、プロセス１１００を開始するとブロック１１１０が実行される。このブロックは初めに、そのドキュメントに関連したＨＴＭＬタグ内にある任意のテキストを含む全テキストをそのドキュメントから抽出する。その後、一度に１文ずつ行なわれる自然言語処理を容易にするために、各ドキュメントのためのテキストが従来の一文切出し処理によって分解され、各文（または疑問文）がファイル内で別個のラインを占めるテキストファイルとなる。その後、ブロック１１１０が（図１３Ａに関連して詳細に後述する）ＮＬＰルーチン１３００をそのドキュメント内のテキストの各ラインごとに別個に呼出して、このドキュメントを分析し、そのラインのための論理形式三つ組の対応の集合を構成してデータセット１０３０内にローカルに記憶する。これらの動作が完了すれば、ブロック１１１０およびプロセス１１００の実行が終了する。

【００７１】

図１０Ａおよび図１０Ｂに示す本発明の具体的な実施例において用いられるような本発明の検索プロセス１２００のフローチャートを図１２Ａおよび図１２Ｂに集合的に示す。図１２Ａおよび図１２Ｂの図面の正しい配列は図１２に示すとおりである。（図６Ａおよび図６Ｂに示し、詳細に前述した）検索プロセス６００とは対照的に、図１２Ａおよび図１２Ｂに示す全動作は共通のコンピュータシ

ステム、ここではＰＣ１０７０（図１０Ａおよび図１０Ｂ参照）において行なわれる。理解を助けるため、以下の説明においては図１０Ａおよび図１０Ｂを同時に参照されたい。

【００７２】

プロセス１２００を開始すると、ブロック１２０５が初めに実行される。このブロックは実行されるとユーザにフルテキストのクエリを入力させる。一旦このクエリが得られると、実行経路は分岐して経路１２０７によってブロック１２１０へ、および経路１２４３によって経路１２４５へ進む。ブロック１２４５は実行されるとＮＬＰルーチン１３５０を呼出してクエリを分析し、対応の論理形式三つ組の集合を構成し、それをローカルにメモリ１０７５内に記憶する。ブロック１２１０は実行されると、破線１２１５で示すようにフルテキストのクエリをサーチエンジン１０９０に送信する。この時点で、サーチエンジンはクエリに応答してブロック１２２０を実行して、ドキュメントレコードの集合とこのようなレコードの各々に関連した関連の論理形式三つ組との両方を検索する。この集合と関連の論理形式三つ組とが検索されれば、両方は破線１２３０で示すようにプロセス１２００に与えられ、具体的にはそこにおけるブロック１２４０に与えられる。ブロック１２４０は単にこの情報をサーチエンジン１０９０から受け、それを後に使用するためにメモリ１０７５内に記憶する。ブロック１２４５における動作をブロック１２１０、１０９０および１２２０における動作と本質的に並列的に行われるものと説明したが、ブロック１２４５における動作は実際の実行上の観点からブロック１２１０、１０９０または１２２０における動作の前または後に直列的に行なわれてもよい。

【００７３】

論理形式三つ組の集合がクエリと検索された各ドキュメントレコードとの両方のためにメモリ１０７５へと記憶されれば、ブロック１２５０が実行される。このブロックは詳細に前述した態様で、クエリ内の論理形式三つ組の各々を、検索された各ドキュメントレコードのための論理形式三つ組の各々と比較して、クエリ内の任意の三つ組と対応の任意のドキュメントの任意の三つ組との間の一致を突き止める。一旦ブロック１２５０が完了すると、ブロック１２５５が実行され

て、一致する三つ組を示さない、すなわち、クエリ内の任意の三つ組と一致する三つ組を有さないドキュメントに対する検索された全レコードを廃棄する。その後、ブロック１２６０が実行される。ブロック１２６０によって、残る全ドキュメントレコードが、前述のように、対応の各ドキュメントごとに存在する一致する三つ組の関係のタイプとそれらの重みとに基づいてスコアを割当てられ、ドキュメントレコードがそのように重み付けされれば、ブロック１２６５が実行されてスコアの降順にレコードをランク付ける。最後に、ブロック１２７０が実行されて、典型的には最も高いスコアを示す予め規定された小グループ、典型的には５または１０のドキュメントレコードについてレコードをランク順に表示する。その後、ユーザはたとえばコンピュータシステム１０７０によって表示されている対応のボタンの上でマウスを適切に「クリックする」ことによって、ランク付けされたドキュメントレコードの次のグループをそのシステムに表示させ、以下同様にユーザがランク付けされた全ドキュメントレコードを順に十分に調べる（そしてその中の関心のある任意のドキュメントにアクセスし、それを調べる）までその動作を行なう。この時点で、プロセス１２００は完了され、実行が終了する。

【００７４】

図１３Ａは、図１１に示す三つ組発生プロセス１１００内で実行されるＮＬＰルーチン１３００のフローチャートを示す。前述のように、ＮＬＰルーチン１３００は索引付けされるべき入来するドキュメント、具体的にはそのためにテキストの１ラインを分析し、そのドキュメントのための論理形式三つ組の対応の集合を構成し、それをローカルに図１０Ａおよび図１０Ｂに示すデータセット内に記憶する。ルーチン１３００は、図７に示し、詳細に前述したＮＬＰルーチン７００と本質的に同様に動作する。

【００７５】

特に、ルーチン１３００が開始されると、ブロック１３１０が最初に実行されて、入力テキストのラインを処理して図５Ａに示す例示のグラフ５１５のような論理形式グラフを生成する。その後、図１３Ａに示すように、ブロック１３２０が実行されてそのグラフから対応の論理形式三つ組の集合を抽出する（読出す）

。一旦これが起こると、ブロック１３３０が実行されて別個に、かつ区別してフォーマット化されたテキスト文字列としてこのような論理形式三つ組の各々を生成する。最後に、ブロック１３４０が実行されて、入力されたテキストのそのラインと、一連のフォーマット化されたテキスト文字列として、そのラインのための論理形式三つ組の集合とがデータセット１０３０に保存される。この集合が完全に記憶されれば、ブロック１３００の実行を終了する。これに替えて、論理形式三つ組ではなく異なる形式、たとえば論理形式グラフまたはそのサブグラフが本発明に関連して用いられるのであれば、ブロック１３２０および１３３０を、フォーマット化された文字列としてその特定の形式を発生するように、そしてブロック１３４０が論理形式三つ組の代わりにその形式をデータセットに記憶するように、容易に変更できるであろう。

【００７６】

図１３Ｂは、検索プロセス１２００内で実行されるＮＬＰルーチン１３５０のフローチャートを示す。前述のように、ＮＬＰルーチン１３５０はユーザ$_j$によって（図１０Ａおよび図１０Ｂに示す）ドキュメント検索アプリケーション１０８５に与えられるクエリを分析し、そのための対応の論理形式三つ組の集合を構成し、メモリ１０７５内にローカルに記憶する。図１３Ａに関連して詳細に前述したルーチン１３００とルーチン１３５０との間の動作上の唯一の違いは、対応の三つ組が記憶される場所である。すなわち、ＮＬＰルーチン１３００におけるブロック１３４０の実行ではデータセット１０３０に記憶され、ＮＬＰルーチン１３５０におけるブロック１３９０の実行ではメモリ１０７５に記憶される、という点である。ルーチン１３５０の他のブロック、具体的にはブロック１３６０、１３７０および１３８０によって行なわれる動作はルーチン１３００のブロック１３１０、１３２０および１３３０によってそれぞれ行われるのと実質的に同じであるので、前者のブロックの詳細な説明をは省略する。

【００７７】

図１Ａに関連して一般的に前述したような本発明の検索プロセスの性能を試験的に試すために、本発明の検索システムにおいてサーチエンジンとしてALTA VISTAサーチエンジンを用いた。インターネット上で誰もがアクセス可能なこのエン

ジンは３１００万を超えるウェブページが索引付けされていると称されている従来の統計的サーチエンジンであり、広く用いられている（概算で現在１日当り２８００万ヒットを記録している。）。本発明の検索プロセス６００を、MICROSOFT OFFICE 97プログラムスイートの一部を成す文法チェッカー内に含まれる、辞書ファイルを含むさまざまな自然言語処理コンポーネントを用いて、一般的なPentium 90 MHzのＰＣ上で実現した（「OFFICE」および「OFFICE 97」はワシントン州レドモンドのMicrosoft Corporationの商標である）。我々はオンラインのパイプライン処理モデルを用いた。すなわち、続く結果をユーザが待っている間にドキュメントがパイプラインの態様でオンラインで集められ、処理された。この特定のＰＣは、各センテンスごとに論理形式三つ組を発生するのに約３分の１秒から約２分の１秒を要した。

【００７８】

サーチエンジンに与えるためのフルテキストのクエリを作るようボランティアに依頼した。合計１２１個の広範囲の互いに異なるクエリが作られ、その代表的なものは「Why was the Celtic civilization so easily conquered by the Romans?（なぜケルト文明はローマ人によって簡単に征服されたのか?）」、「Why do antibiotics work on colds but not on viruses?（抗生物質はなぜ風邪に効くのにウィルスには効かないのか?）」、「Who is the governor of Washington?（ワシントン州の知事は誰か?）」、「Where does the Nile cross the equator?（ナイル川が赤道と交差するのはどこか?）」、および「When did they start vaccinating for small pox?（天然痘の予防接種が始められたのはいつ頃か?）」といったものであった。これらの１２１個の各クエリをALTA VISTAサーチエンジンに与え、各クエリに応答して戻った利用可能なものからなる上位３０のドキュメントを得た。クエリの中には３０未満のドキュメントしか戻らないものがあり、その場合は戻った全ドキュメントを使用した。全１２１クエリに対して、延べ３３６１ドキュメント（すなわち、「生の」ドキュメント）が得られた。

【００７９】

３３６１のドキュメントと１２１のクエリとの各々が本発明のプロセスによって分析されて論理形式三つ組の対応の集合が生成された。集合は適切に比較され

、結果のドキュメントが前述のように選択され、スコアおよびランクをつけられた。

【００８０】

３３６１のドキュメントの全てを、そのドキュメントが得られた対応のクエリとの関連性について手作業で別個に評価した。関連性を評価するため、発明者の具体的な実験目的を知らない一人の人を評価者として利用し、これらの３３６１ドキュメントの各々をその対応のクエリとの関連について「最適」、「関連あり」または「関連なし」として手作業で主観的にランク付けした。最適なドキュメントとは、対応のクエリに対する明らかな回答を含むものであるとされた。関連のあるドキュメントとは、クエリに対する明らかな回答を含まないがそれにもかかわらず関連性のあるもののことであるとされた。関連のないドキュメントとは、クエリに対する有益な回答ではないもの、すなわち、クエリに関連がないか、英語以外の言語によるか、またはALTA VISTAエンジン（すなわち、「cobweb」リンク）によって与えられた対応のＵＲＬからは検索できないドキュメントのことであるとされた。評価精度を高めるため、二人目の評価者がこれらの３３６１ドキュメント内のサブセット、具体的には、その対応のクエリにおける論理形式三つ組と一致する少なくとも１つの論理形式三つ組を示したドキュメント（３３６１ドキュメントのうち４３１）と、それまでに関連ありまたは最適であるとランク付けされたが一致する論理形式三つ組を有さないドキュメント（３３６１ドキュメントのうち１０２）とを調査した。ドキュメントに対するこれらのランク付けの意見の相違があれば、それは「仲裁者」となる三人目の評価者によって検討された。

【００８１】

この実験の結果として、関連した全ドキュメントにおいて、本発明の検索システムは、ALTA VISTAサーチエンジンが戻す生のドキュメントよりも改善を示したことが観察された。全体の（すなわち、選択された全ドキュメントの）精度では約１６％から約４７％へと２００％程度改善し、上位５ドキュメント内では約２６％から約５１％へと約１００％改善した。加えて、本発明のシステムの使用によって、最適であるとして戻された最初のドキュメントの精度は生のドキュメン

トのそれに対して約１７％から約３５％へと約１１３％改善された。

【００８２】

本発明を統計的サーチエンジンでの使用を例として具体的に説明したが、本発明はそれに限定されない。その点では、本発明は実質的にいかなるタイプのサーチエンジンによって得られた検索ドキュメントをも処理してそのエンジンの精度を高めるよう用いることができる。

【００８３】

論理形式三つ組内の各種の属性ごとに固定した重みを用いる代わりに、これらの重みを動的に変化させ、実際のところ、適応的としてもよい。これを達成するため、たとえばベイズのネットワークまたはニューラルネットワークのような学習メカニズムを本発明のプロセスに適切に組込み、各種の論理形式三つ組のための重み数値を学習経験に基づいて最適な値に変えてもよい。

【００８４】

本発明のプロセスは論理形式三つ組が正確に一致することを必要としたが、三つ組の間で十分に類似した意味内容を識別する目的のために、一致判断の基準を緩和してパラフレーズも一致とみなすようにしてもよい。パラフレーズは語彙的であってもよく構造的であってもよい。語彙的パラフレーズの例は上位語または類義語であろう。構造的パラフレーズの例は同格関係にある名詞または関係節の使用である。たとえば、「大統領、ビル・クリントン」のような同格関係にある名詞構成は「大統領であるビル・クリントン」のような関係節構成と一致すものとみなされるべきである。意味レベルでは、２つの語が互いにいかに意味的に類似しているかについてきめの細かい判断をすることができ、それによって、「どこでコーヒーが栽培されるか」というクエリと「コーヒーは熱帯山岳地方でよく栽培される」というようなコーパス（文例）文との間の一致を認めることができる。加えて、一致が存在するか否かを判断するための手順を、与えられたクエリのタイプによって変更することができる。たとえば、何かが存在する場所についてクエリが尋ねていれば、その手順は、ある文がクエリと一致するとみなされるためには、テストされている文と関連した任意の三つ組内に「場所」属性が存在していなければならない、と要求するであろう。したがって、論理形式三つ組の

「一致」とは同一の一致だけをいうのみならず、このような緩和された、判断を含むような、変更された一致条件の全てから生じるものをも包含するように包括的に規定される。

【００８５】

さらに、本発明をグラフィックス、表、映像またはその他のような非テキスト情報の検索を中心とする他の処理技術と容易に組合せて全体の精度を高めることができる。一般に、ドキュメント中の非テキスト内容にはよく、たとえば図の記号または短い説明のような、そのドキュメント中の言語的（テキストによる）描写が付随するものである。したがって、本発明のプロセス、特にその自然言語成分の使用を、非テキスト内容にしばしば付随する言語的描写を分析し、処理するために用いることができる。クエリに意味的に関連した言語的内容を示すドキュメントの集合を初めに探し、次に、このドキュメントの集合をそれらの非テキスト内容に関して処理することによって、本発明の自然言語処理技術を用いて関連あるテキスト内容および非テキスト内容を有するドキュメントを検索することができる。これに替えて、ドキュメント検索を初めに非テキスト内容について行なってドキュメントの集合を取出し、次に本発明の技術によってそのドキュメントの集合をそれらの言語的内容について処理することで関連のあるドキュメントを検索してもよい。

【００８６】

本発明の教示を採用したさまざまな実施例を図示し、詳細に説明したが、当業者であればこれらの教示をなお利用する多くの他の実施例に容易に想到することができるであろう。

【図面の簡単な説明】

【図１】　我々の本発明に従う情報検索システム５の非常に高いレベルのブロック図である。

【図２】　我々の本発明の教示を利用する、図１に示されるタイプの情報検索システム２００の高いレベルの実施例を示す図である。

【図３】　図２に示されるシステム２００内に含まれる、特定的にはクライアントパーソナルコンピュータであるコンピュータシステム３００を示すブロッ

ク図である。

【図４】　図３に示されるコンピュータ３００内で実行されるアプリケーションプログラム４００を示す非常に高いレベルのブロック図である。

【図５】　ＡからＤは、種々の複雑さを有する英語の文の種々の対応例、およびそれらに関する対応の論理形式要素を示す図である。

【図６】　図６は図６Ａおよび図６Ｂの図面の正しい配置を示す図であり、図６Ａおよび図６Ｂは、我々の発明の検索プロセス６００のフローチャートを合わせて示す図である。

【図７】　プロセス６００内で実行されるＮＬＰルーチン７００のフローチャートを示す図である。

【図８Ａ】　例として、一致する論理形式三つ組の重み付け表８００を示す図である。

【図８Ｂ】　例示的な質問および統計的に検索された例の３つのドキュメントの組に関する、図６Ａおよび図６Ｂに全て示されているブロック６５０、６６０、６６５および６７０内で行なわれる、我々の発明の教示に従う論理形式三つ組の比較、ドキュメントのスコア付け、ランク付けおよび選択処理を視覚的に示す図である。

【図９】　ＡからＣは、我々の本発明の教示を採用した情報検索システムの３つの異なった実施例をそれぞれ示す図であり、Ｄは我々の本発明のさらに別の異なった実施例を実現するにあたり使用される、図９Ｃに示されるリモートコンピュータ（サーバ）９３０の代替的な実施例を示す図である。

【図１０】　図１０は図１０Ａおよび図１０Ｂの図面の正しい配置を示す図であり、図１０Ａおよび図１０Ｂは我々の本発明のさらに別の実施例であって、各ドキュメントに関する論理形式三つ組が、それらに関するドキュメントレコードとともに予め計算されて記憶され、後のドキュメント検索動作時にアクセスされるものを合わせて示す図である。

【図１１】　図１０Ａおよび図１０Ｂに示されるドキュメント索引付けエンジン１０１５によって行なわれる三つ組生成処理１１００を示す図である。

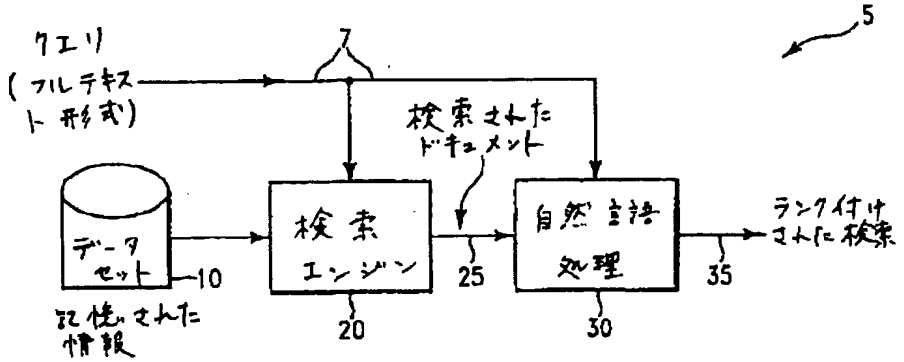
【図１２】　図１２は図１２Ａおよび図１２Ｂの図面の正しい配置を示す図

であり、図１２Ａおよび図１２Ｂは、図１０Ａおよび図１０Ｂに示されるコンピュータシステム３００内で実行される我々の発明の検索処理１２００のフローチャートを合わせて示す図である。

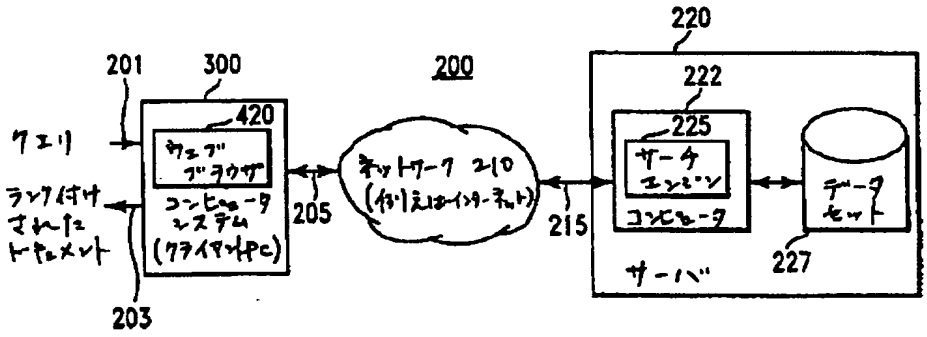
【図１３Ａ】　三つ組生成処理１１００内で実行されるＮＬＰルーチン１３００のフローチャートを示す図である。

【図１３Ｂ】　検索処理１２００内で実行されるＮＬＰルーチン１３５０のフローチャートを示す図である。

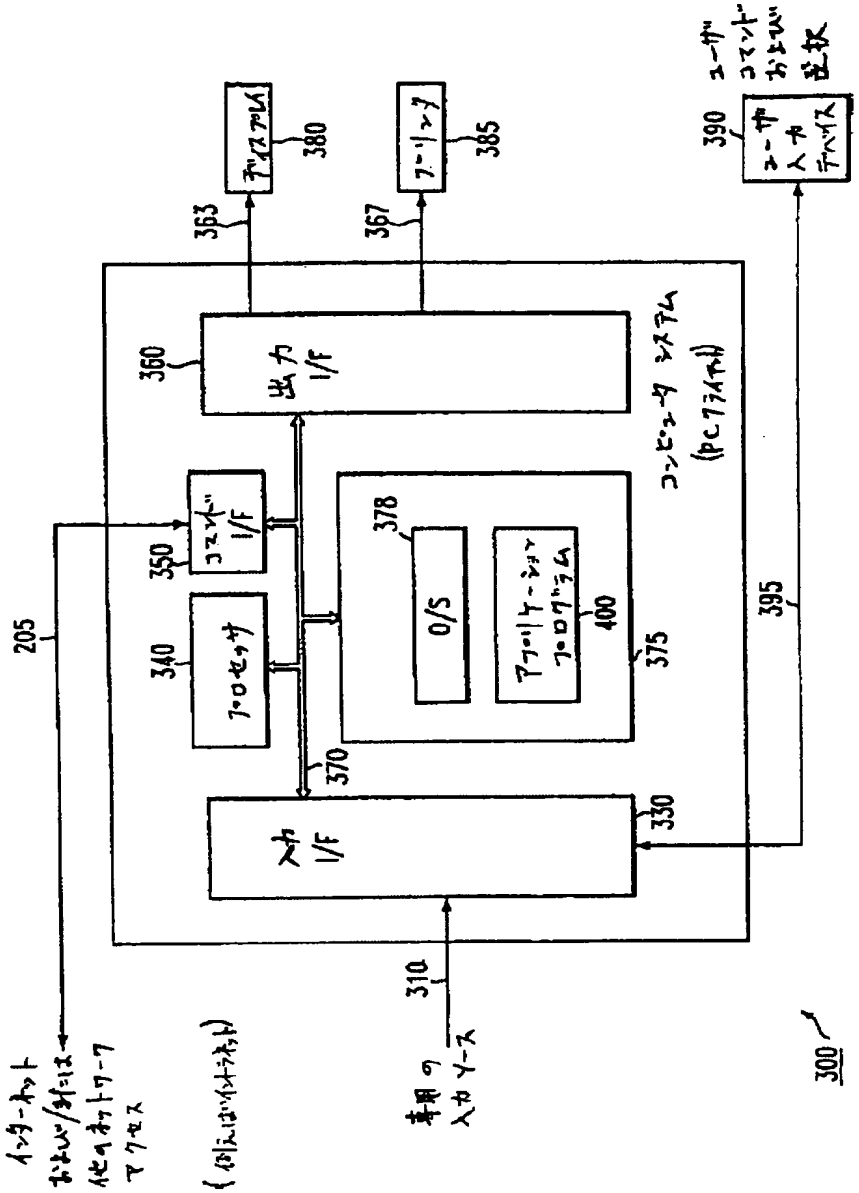
【図１】



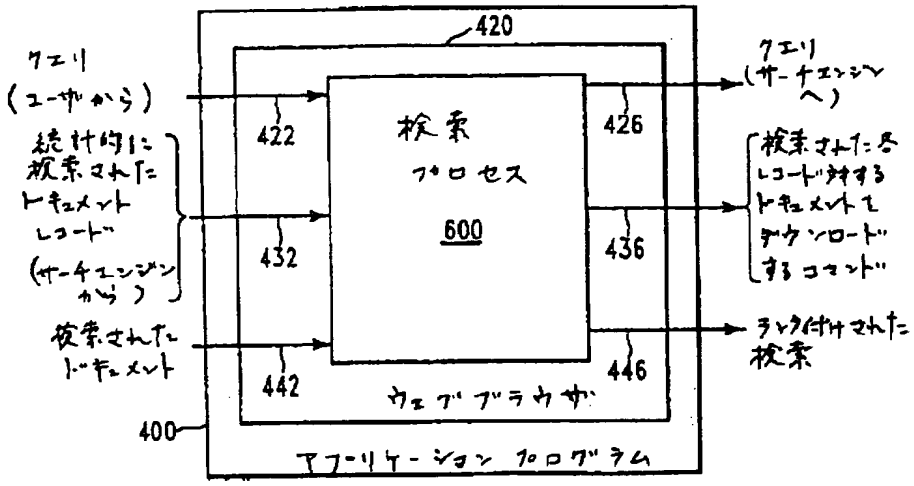
【図２】



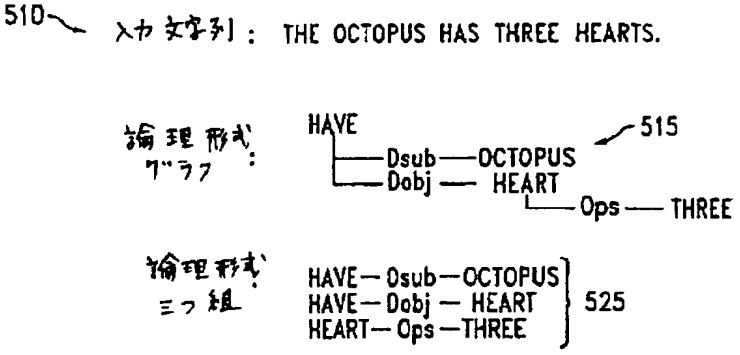
【図３】


ユーザ コマンド および 応答
390
ユーザ 入力 デバイス
ディスプレイ
380
プリンタ
385
363
367
360
出力 I/F
コンピュータ システム (PC アプリケーション)
コマンド I/F
350
340
プロセッサ
378
O/S
アプリケーション プログラム
400
375
370
入力 I/F
330
395
205
310
インターネット および/または その他のネットワーク アクセス
(例えばパケット交換)
車両の 入力 ソース
300

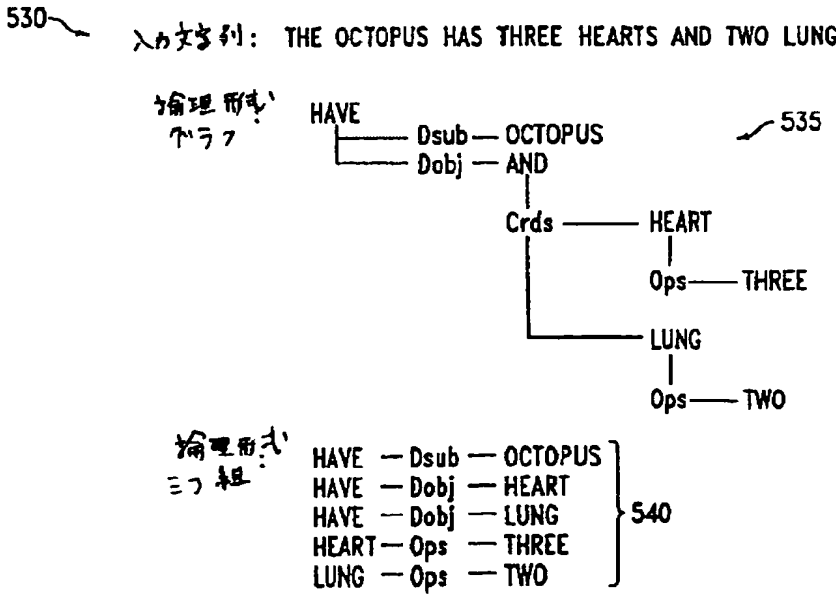
【図４】


420
クエリ
(ユーザから)
422
検索
プロセス
600
426
クエリ
(サーチエンジンへ)
統計的に
検索された
ドキュメント
レコード
(サーチエンジンから)
432
436
検索された
ドキュメント
442
446
ランク付けされた
検索
ウェブブラウザ
400
アプリケーション プログラム

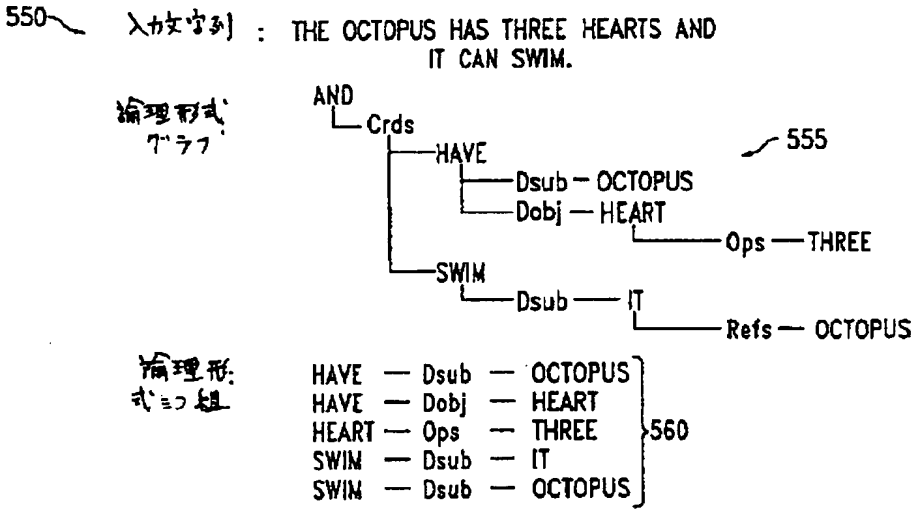
【図５Ａ】


510
入力文字列： THE OCTOPUS HAS THREE HEARTS.
論理形式
グラフ：
HAVE
Dsub — OCTOPUS
Dobj — HEART
Ops — THREE
515
論理形式
三つ組：
HAVE — Dsub — OCTOPUS
HAVE — Dobj — HEART
HEART — Ops — THREE
525

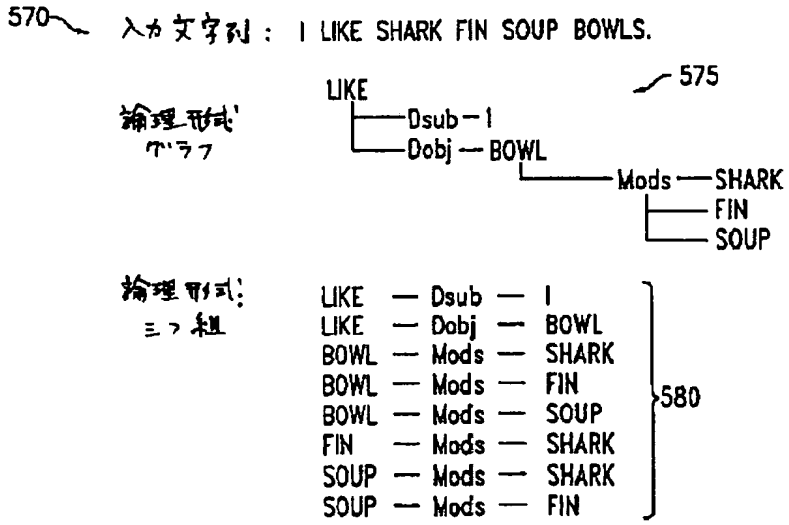
【図５Ｂ】


530
入力文字列： THE OCTOPUS HAS THREE HEARTS AND TWO LUNGS.
論理形式
グラフ
HAVE
Dsub — OCTOPUS
Dobj — AND
Crds — HEART
Ops — THREE
LUNG
Ops — TWO
535
論理形式
三つ組：
HAVE — Dsub — OCTOPUS
HAVE — Dobj — HEART
HAVE — Dobj — LUNG
HEART — Ops — THREE
LUNG — Ops — TWO
540

【図５Ｃ】



【図５Ｄ】



【図６】

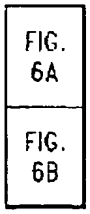
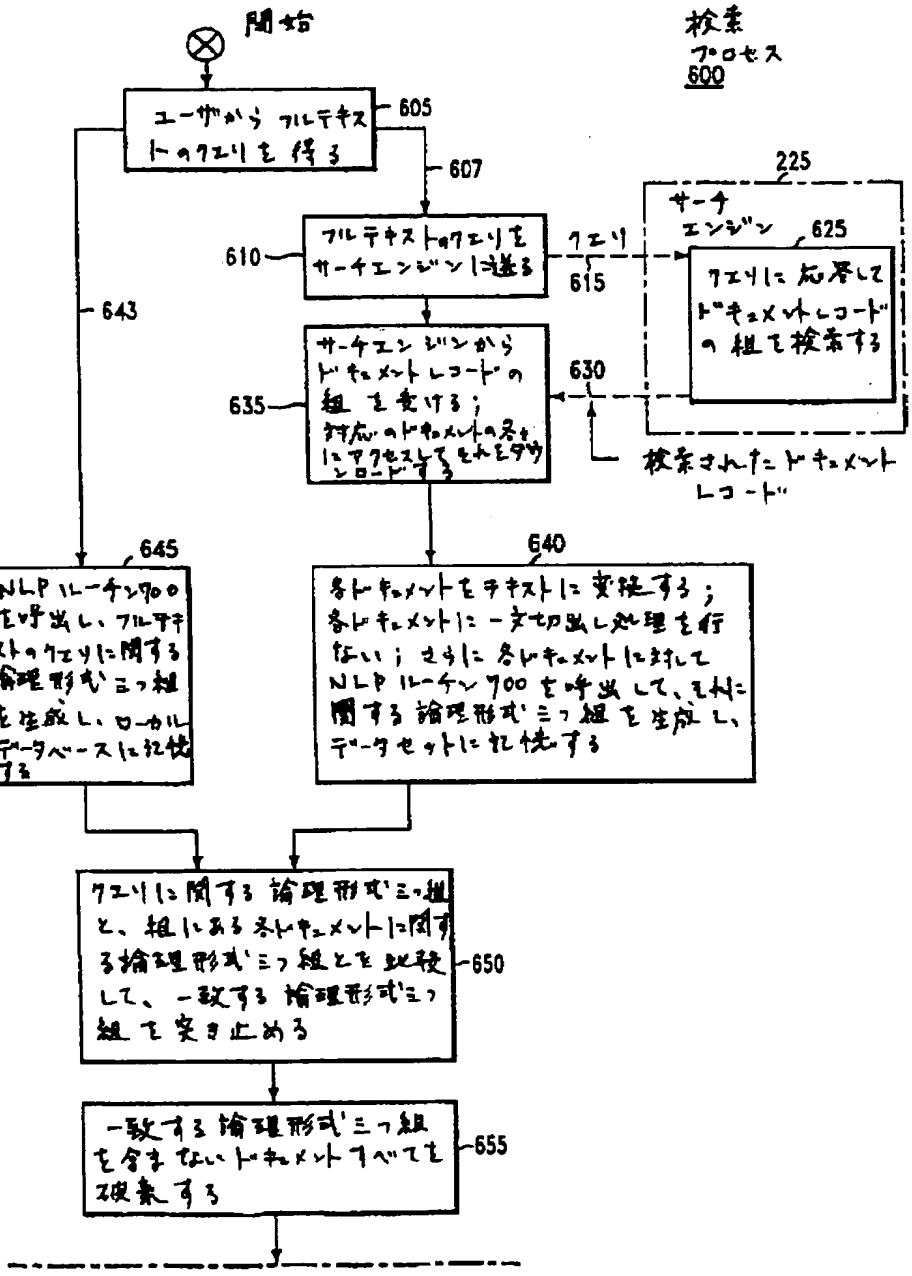
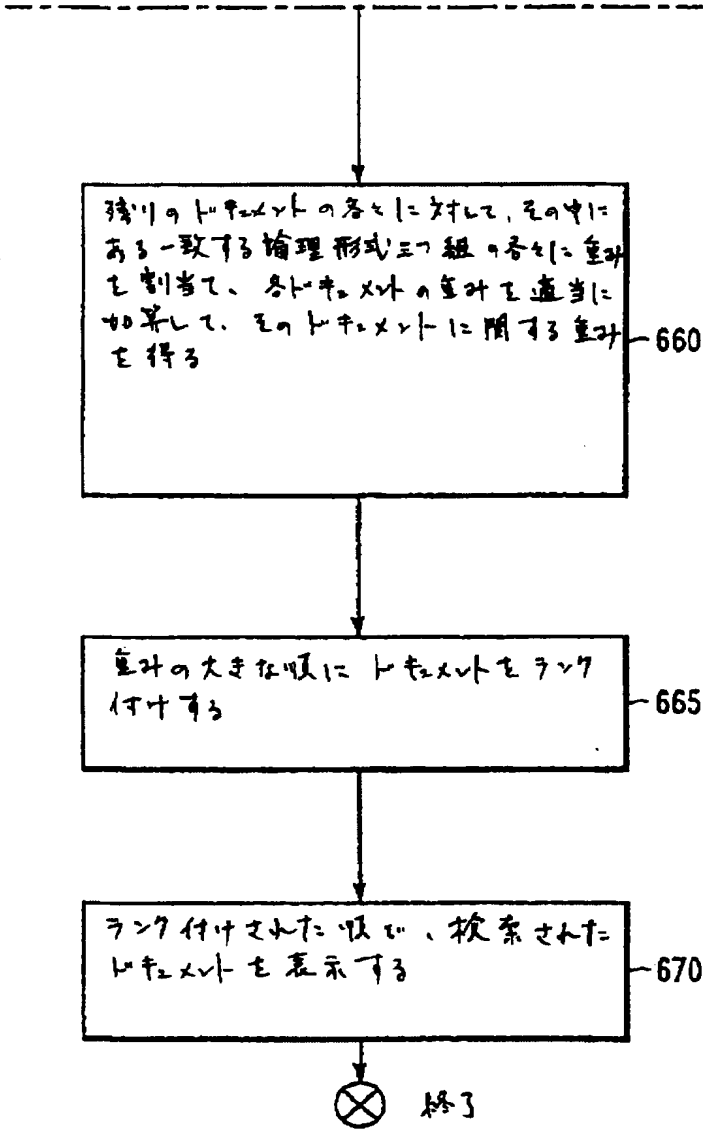


FIG. 6

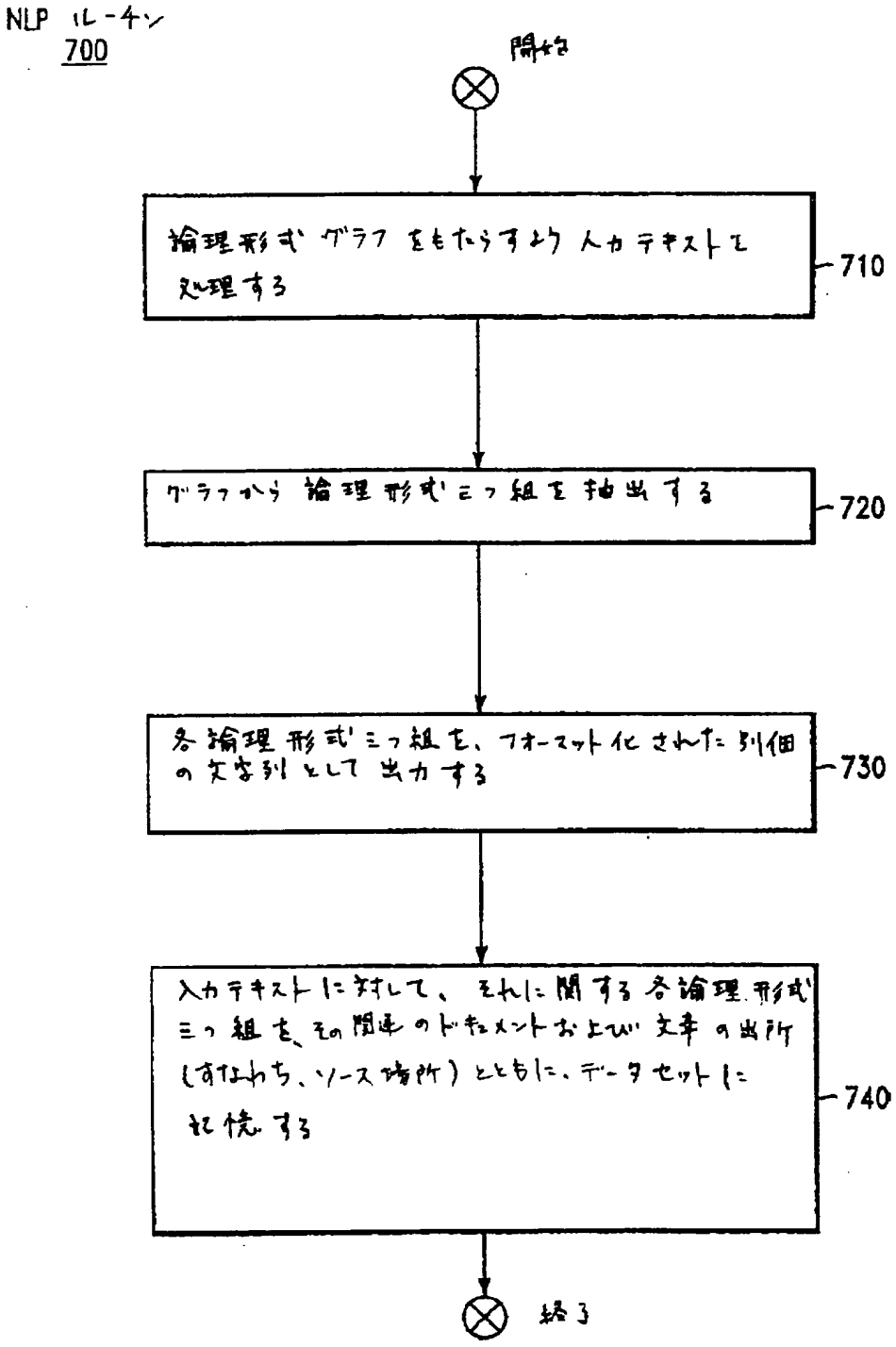
【図６Ａ】


開始
検索
プロセス
600
ユーザからフルテキストのクエリを得る
605
607
225
サーチエンジン
625
フルテキストのクエリをサーチエンジンに送る
610
クエリ
615
クエリに応答してドキュメントレコードの組を検索する
643
サーチエンジンからドキュメントレコードの組を受ける；対応のドキュメントの各々にアクセスしてそれをダウンロードする
635
630
検索されたドキュメントレコード
645
640
NLPルーチン700を呼出し、フルテキストのクエリに関する論理形式三つ組を生成し、ローカルデータベースに記憶する
各ドキュメントをテキストに変換する；各ドキュメントに一文切出し処理を行ない；さらに各ドキュメントに対してNLPルーチン700を呼出して、それに関する論理形式三つ組を生成し、データセットに記憶する
クエリに関する論理形式三つ組と、組にある各ドキュメントに関する論理形式三つ組とを比較して、一致する論理形式三つ組を突き止める
650
一致する論理形式三つ組を含まないドキュメントすべてを破棄する
655

【図６Ｂ】


残りのドキュメントの各々に対して、その中にある一致する論理形式三つ組の各々に重みを割当て、各ドキュメントの重みを適当に加算して、そのドキュメントに関する重みを得る
660
重みの大きな順にドキュメントをランク付けする
665
ランク付けされた順で、検索されたドキュメントを表示する
670
終了

【図７】

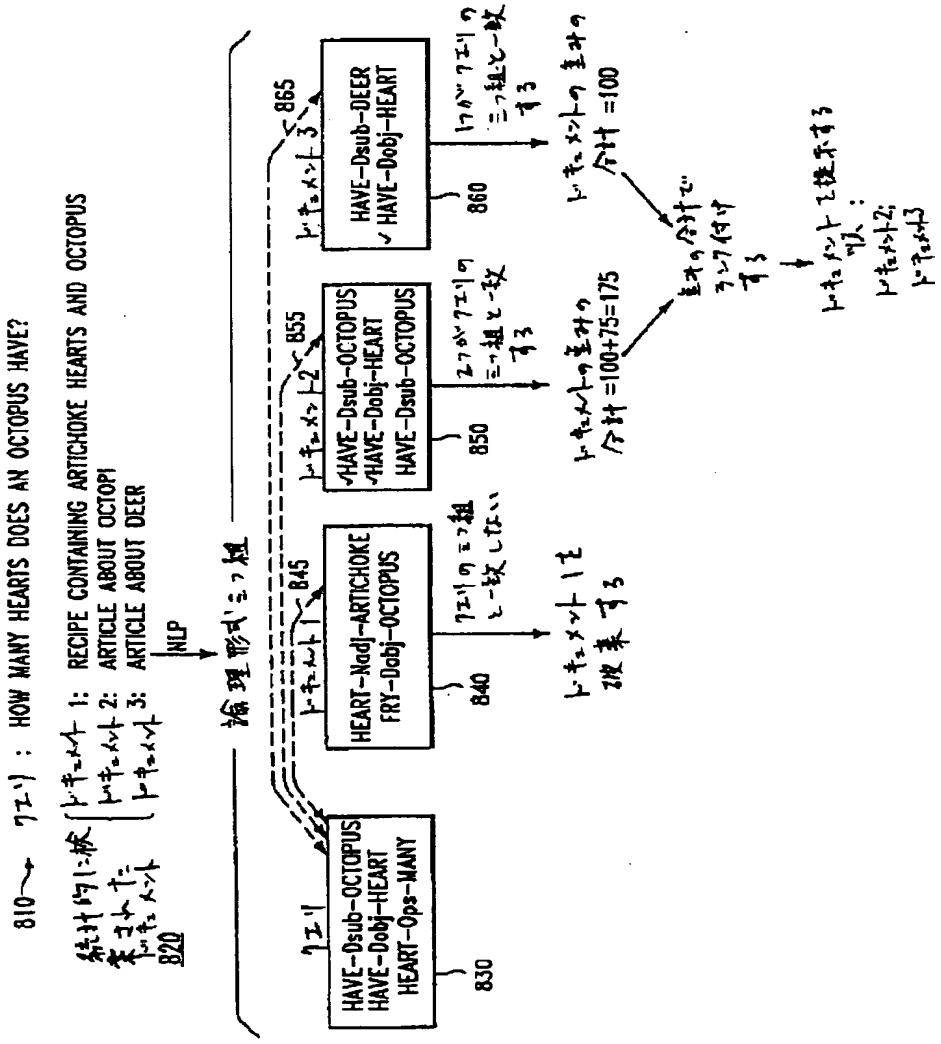

NLP ルーチン
700
開始
論理形式グラフをもたらすよう入力テキストを処理する
710
グラフから論理形式三つ組を抽出する
720
各論理形式三つ組を、フォーマット化された別個の文字列として出力する
730
入力テキストに対して、それに関する各論理形式三つ組を、その関連のドキュメントおよび文章の出所（すなわち、ソース場所）とともに、データセットに収集する
740
終了

【図８Ａ】

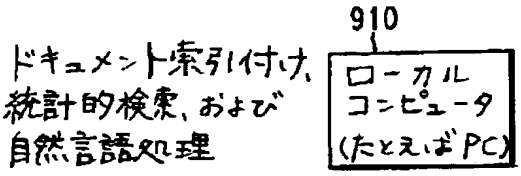
一致するﾐﾆ組の関係子のタイプ	重み
Dobj	100
Dsub	75
Ops	10
Nodj	10

800 一致論理形式ﾐﾆ組重み付け表

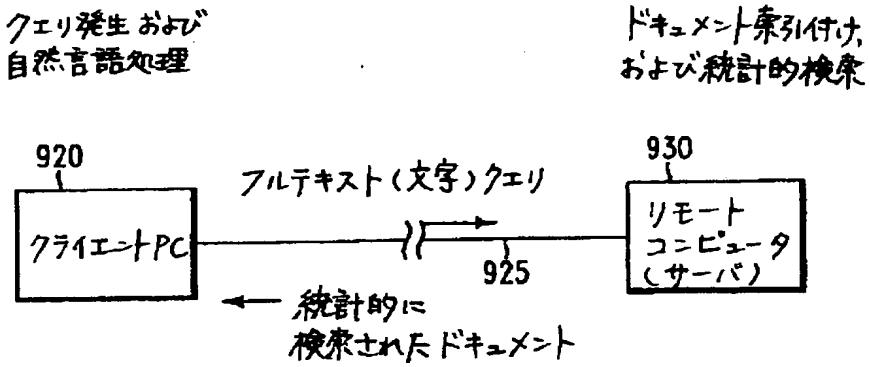
【図８Ｂ】



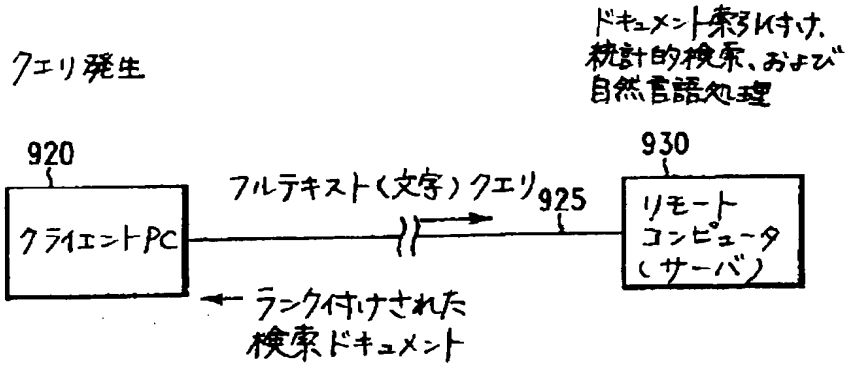
【図９Ａ】



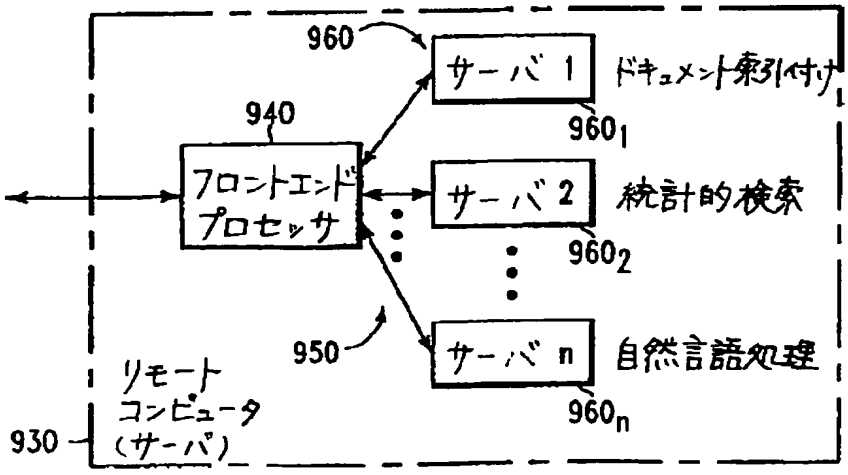
【図９Ｂ】



【図９Ｃ】

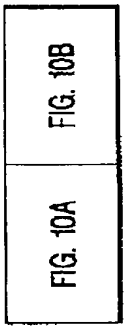


【図９Ｄ】

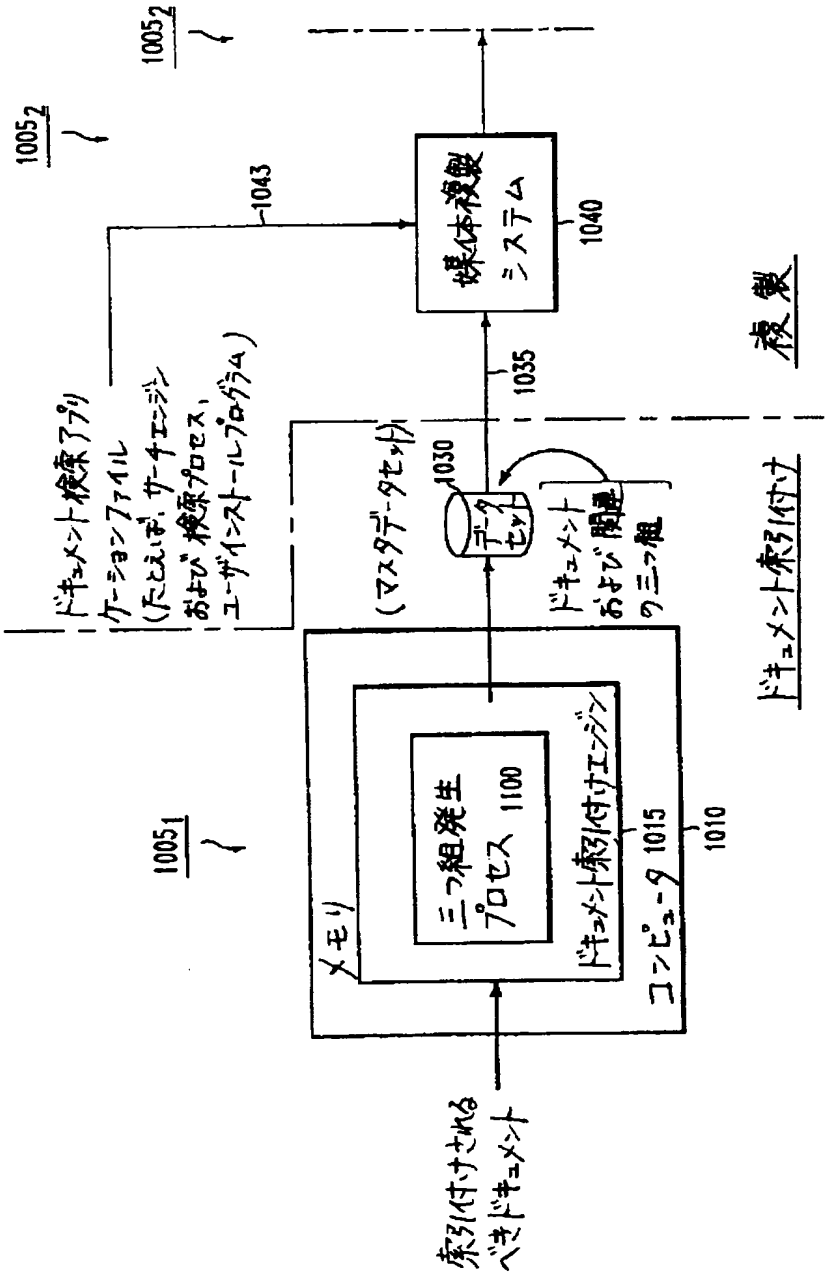


【図１０】

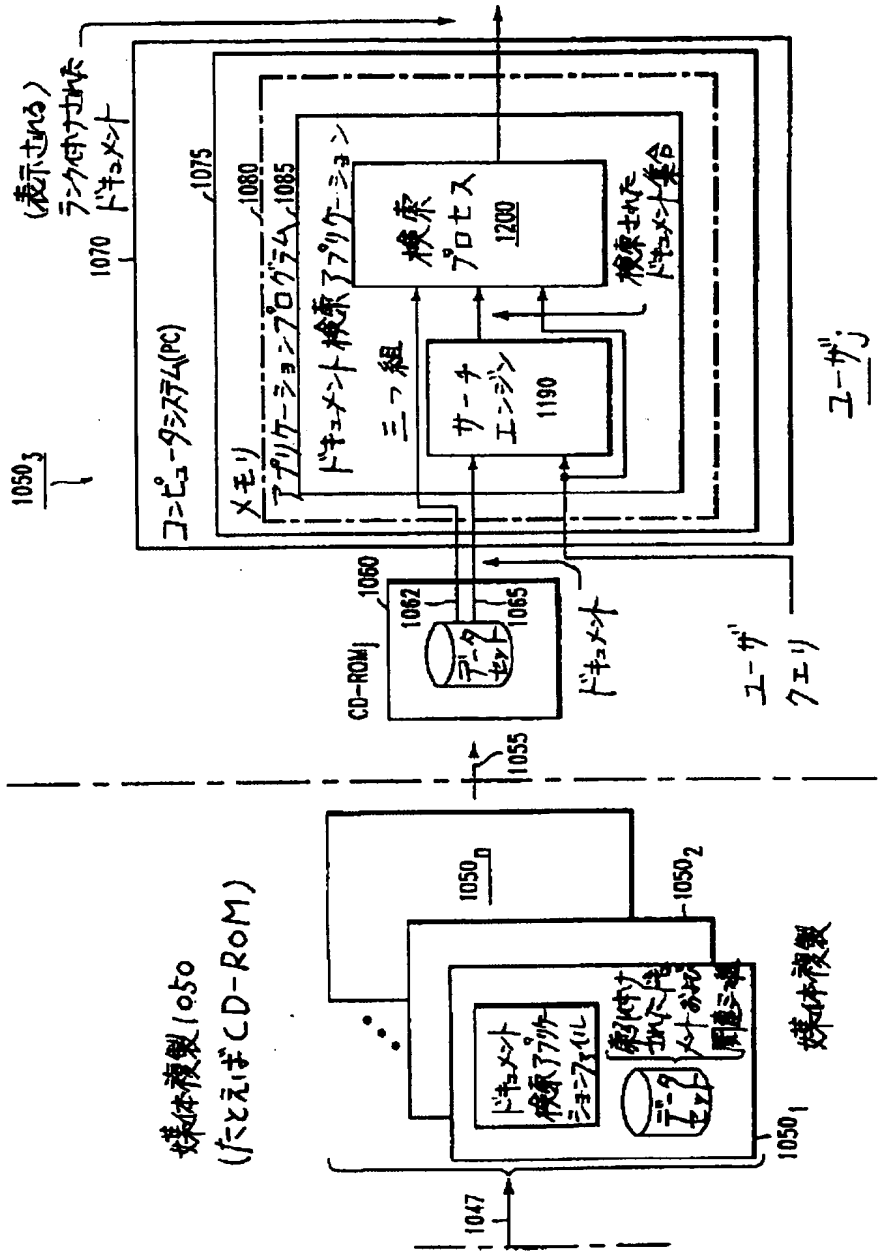
FIG. 10



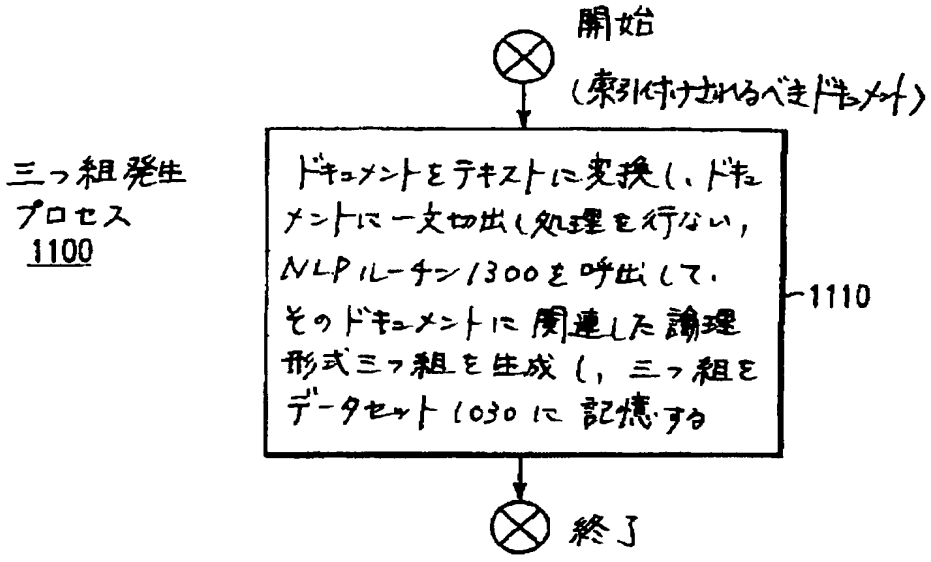
【図１０Ａ】


1005₂
1005₂
1043
媒体複製システム
1040
ドキュメント検索アプリケーションファイル
(たとえば、サーチエンジンおよび検索プロセス、ユーザインストールプログラム)
1035
複製
(マスタデータセット)
1030
データセット
ドキュメントおよび関連のミラー組
ドキュメント索引付け
1005₁
メモリ
三つ組発生プロセス 1100
ドキュメント索引付けエンジン
コンピュータ 1015
1010
索引付けされるべきドキュメント

【図１０Ｂ】


(表示される)
ランク付けされた
ドキュメント
1070
1075
1080
1085
メモリ
アプリケーションプログラム
ドキュメント検索アプリケーション
三つ組
検索
プロセス
1200
サーチ
エンジン
1190
検索された
ドキュメント集合
1050₃
コンピュータシステム(PC)
ユーザ
1060
1062
1065
CD-ROM
ドキュメント
ユーザ
クエリ
1055
媒体複製1050
(たとえばCD-ROM)
1050ₙ
1050₂
ドキュメント
検索アプリケーションファイル
1050₁
媒体複製
1047

【図１１】



【図１２】

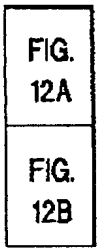
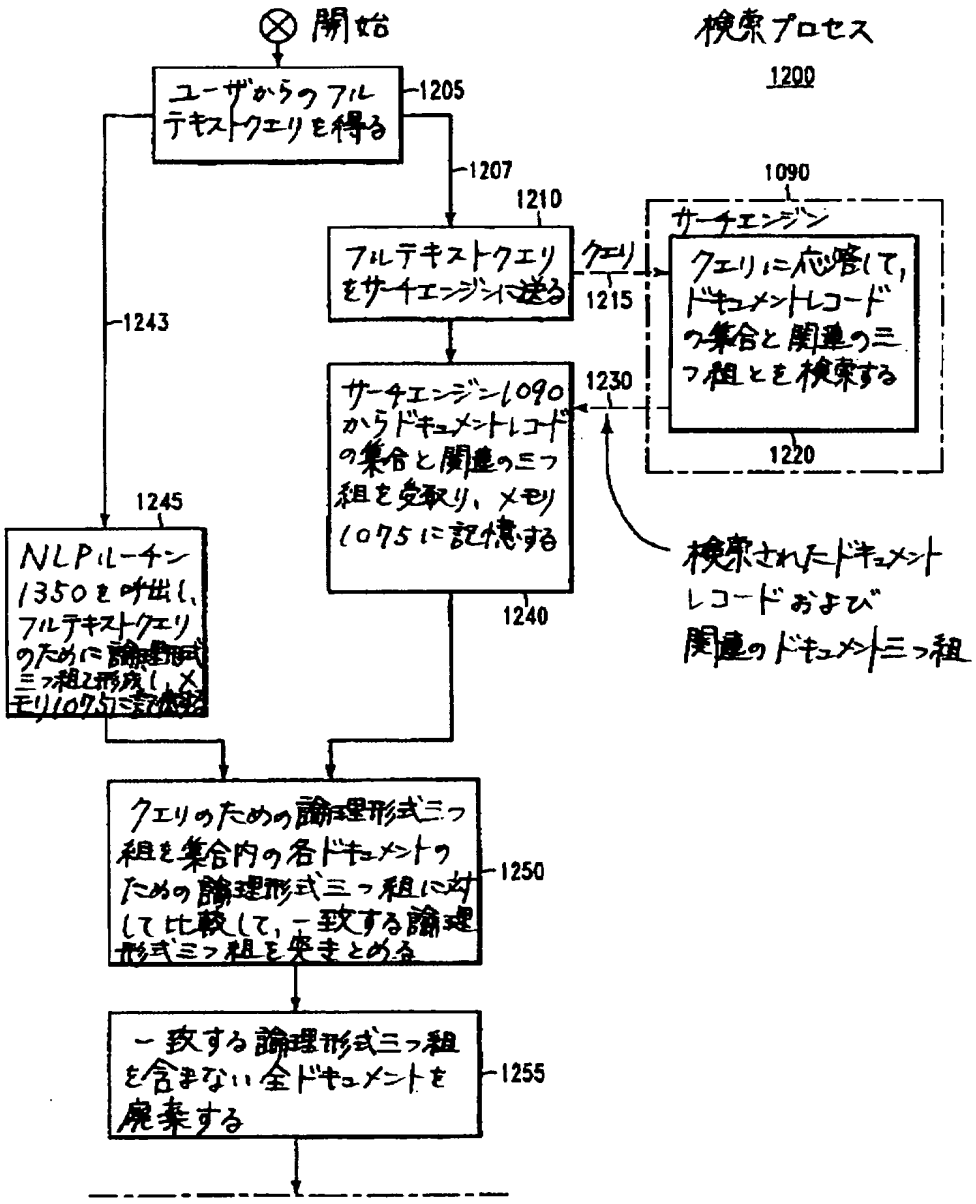
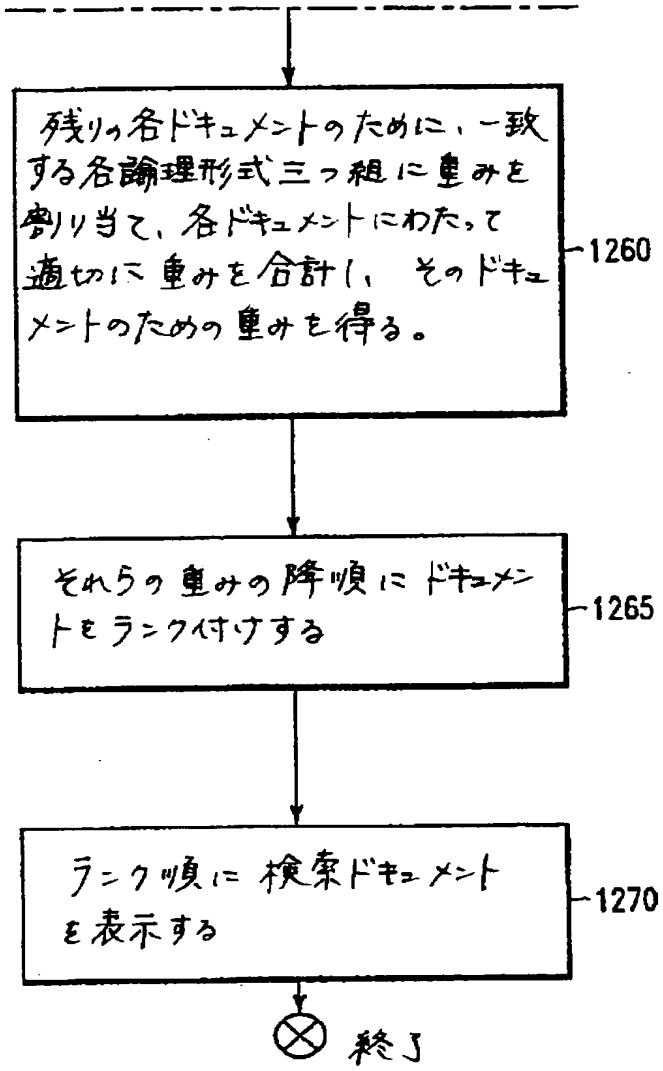


FIG. 12

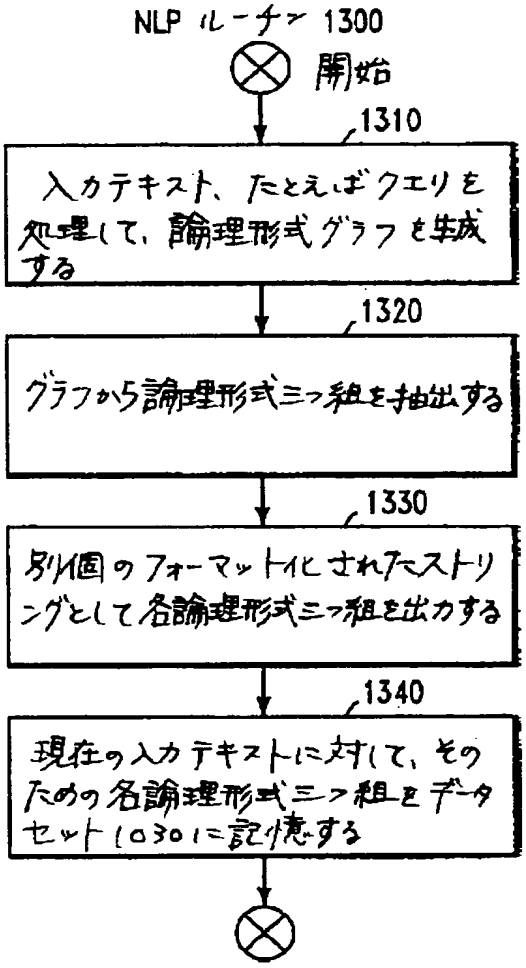
【図１２Ａ】


開始
検索プロセス
1200
ユーザからのフルテキストクエリを得る
1205
1207
1210
フルテキストクエリをサーチエンジンに送る
クエリ
1215
1090
サーチエンジン
クエリに応答して、ドキュメントレコードの集合と関連の三つ組とを検索する
1220
1230
サーチエンジン1090からドキュメントレコードの集合と関連の三つ組を受取り、メモリ1075に記憶する
1240
検索されたドキュメントレコードおよび関連のドキュメント三つ組
1243
1245
NLPルーチン1350を呼出し、フルテキストクエリのために論理形式三つ組を形成し、メモリ1075に記憶する
クエリのための論理形式三つ組を集合内の各ドキュメントのための論理形式三つ組に対して比較して、一致する論理形式三つ組を集まとめる
1250
一致する論理形式三つ組を含まない全ドキュメントを廃棄する
1255

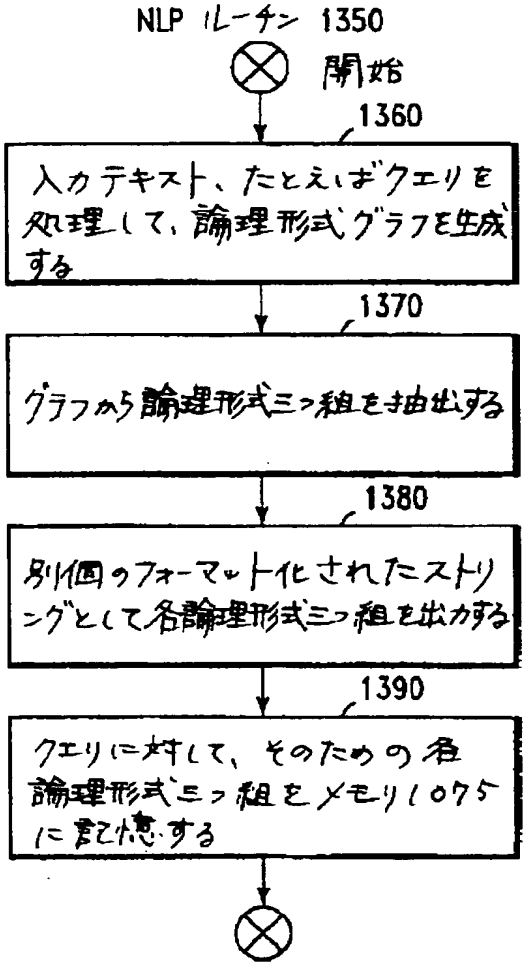
【図１２Ｂ】


残りの各ドキュメントのために、一致する各論理形式三つ組に重みを割り当て、各ドキュメントにわたって適切に重みを合計し、そのドキュメントのための重みを得る。
1260
それらの重みの降順にドキュメントをランク付けする
1265
ランク順に検索ドキュメントを表示する
1270
終了

【図１３Ａ】


NLP ルーチン 1300
開始
1310
入力テキスト、たとえばクエリを処理して、論理形式グラフを生成する
1320
グラフから論理形式三つ組を抽出する
1330
別個のフォーマット化されたストリングとして各論理形式三つ組を出力する
1340
現在の入力テキストに対して、そのための各論理形式三つ組をデータセット1030に記憶する

【図１３Ｂ】


NLP ルーチン 1350
開始
1360
入力テキスト、たとえばクエリを処理して、論理形式グラフを生成する
1370
グラフから論理形式三つ組を抽出する
1380
別個のフォーマット化されたストリングとして各論理形式三つ組を出力する
1390
クエリに対して、そのための各論理形式三つ組をメモリ1075に記憶する

【手続補正書】
【提出日】平成１２年１月２５日（２０００．１．２５）
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】特許請求の範囲
【補正方法】変更
【補正内容】
【特許請求の範囲】
【請求項１】　記憶されているドキュメントをリポジトリ（１０）から検索するための情報検索システム（５）において用いるための装置であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントをリポジトリから検索し、出力ドキュメント集合を規定するための検索システム（２０）を有し、前記装置は、
プロセッサ（３４０）と、
実行可能な命令（６００）が記憶されているメモリ（３７５）とを含み、
プロセッサはメモリに記憶されている命令に応答して、
クエリに応答してそのための第１の論理形式を生じ、第１の論理形式はクエリに関連した語の間の意味的関係を示し、
出力ドキュメント集合内のドキュメントの各別の１つに対して、対応する第２の論理形式を取得し、第２の論理形式は前記１つのドキュメント内の句に関連した語の間の意味的関係を示し、
クエリの第１の論理形式と、出力ドキュメント集合内の複数のドキュメントの各1つのための第２の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定し、
出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力（４４６）として与える、装置。
【請求項２】　各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の１つのドキュメントに関連したレコードである、請求項１に記載の装置。

【請求項３】　クエリのための第１の論理形式と出力ドキュメント集合内の各別のドキュメントのための第２の論理形式との各々はそれぞれ、論理形式グラフ（５１５，５３５，５５５，５７５）、そのサブグラフ、または論理形式三つ組のリスト（５２５，５４０，５６０，５８０）である、請求項２に記載の装置。

【請求項４】　プロセッサは記憶されている命令に応答して、

出力ドキュメント集合内のドキュメントの前記各別の1つのために、記憶媒体から対応の第２の論理形式を読出すか、または

出力ドキュメント集合内の前記各別の１つのドキュメントを分析することによって前記対応の第２の論理形式を生成する、請求項３に記載の装置。

【請求項５】　前記関数は、前記ドキュメントの１つのために、クエリに関連した前記第１の論理形式と前記１つのドキュメントに関連した前記第２の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、プロセッサは記憶されている命令に応答して、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定する、請求項４に記載の装置。

【請求項６】　クエリに関連した前記第１の論理形式と出力ドキュメント集合内の任意のドキュメントに関連した前記第２の論理形式の任意のものとの間の前記一致は同一の一致である、請求項５に記載の装置。

【請求項７】　ユーザからのクエリ（２０１）を取得し、出力ドキュメント集合内の複数のドキュメント（２０３）を前記ランク順に表示するためのクライエントコンピュータ（３００）と、

ネットワーク接続（２０５，２１０，２１５）を介してクライエントコンピュータに接続されるサーバ（２２０）とをさらに含み、前記サーバは前記プロセッサ（３４０）および前記メモリ（３７５）を含み、プロセッサはメモリに記憶されている命令（６００）に応答して、

クライエントコンピュータからクエリを取得し、

出力ドキュメントの集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与える、請求項５に記載の装置。

【請求項８】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項５または６に記載の装置。

【請求項９】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項５または６に記載の装置。

【請求項１０】　前記関数は、クエリに関連した論理形式三つ組の少なくとも１つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的関係のタイプによって定義される、請求項８に記載の装置。

【請求項１１】　プロセッサはメモリに記憶されている命令に応答して、

クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定し、

関連した少なくとも１つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各1つのために、前記一致する論理形式三つ組の各々に関連した意味的関係によって予め規定される重み数値を用いて前記各1つのドキュメント内の一致する論理形式三つ組に重み付けして、前記１つのドキュメントのための１つ以上の重みを形成し、

前記１つ以上の重みの関数として前記１つのドキュメントのためのスコアを計算し、

前記ドキュメントの各１つをその前記スコアに従ってランク付けしてランク順を規定する、請求項１０に記載の装置。

【請求項１２】　クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項８に記載の装置。

【請求項１３】　クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項８に記載の装置。

【請求項１４】　サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリ（１０）から記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各１つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項５、７または８に記載の装置。

【請求項１５】　記憶されているドキュメントをリポジトリ（１０）から検索するための情報検索システム（５）において用いるための方法（６００；１２００）であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントをリポジトリから検索し、出力ドキュメント集合を規定するための検索システム（２０）を有し、前記方法は、

クエリに応答してそのための第１の論理形式を生じる（６４５；１２４５）ステップを含み、第１の論理形式はクエリに関連した語の間の意味的関係を示し、

出力ドキュメント集合内のドキュメントの各別の１つに対して、対応する第２の論理形式を取得する（６４０；１２４０）ステップを含み、第２の論理形式は前記１つのドキュメント内の句に関連した語の間の意味的関係を示し、

クエリの第１の論理形式と、出力ドキュメント集合内の複数のドキュメントの各１つのための第２の論理形式との予め定義された関数（６５０，６５５，６６０；１２５０，１２５５，１２６０）として、出力ドキュメント集合内の複数の

ドキュメントをランク付けして（６６５；１２６５）ランク順を規定するステップと、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与える（６７０；１２７０）ステップとを含む、方法。

【請求項１６】　各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の１つのドキュメントに関連したレコードである、請求項１５に記載の方法。

【請求項１７】　クエリのための第１の論理形式と出力ドキュメント集合内の各別のドキュメントのための第２の論理形式との各々はそれぞれ、論理形式グラフ（５１５，５３５，５５５，５７５）、そのサブグラフ、または論理形式三つ組のリスト（５２５，５４０，５６０，５８０）である、請求項１６に記載の方法。

【請求項１８】　前記取得するステップは、

出力ドキュメント集合内のドキュメントの前記各別の１つのために、記憶媒体から対応の第２の論理形式を読出す（１２４０）か、または

出力ドキュメント集合内の前記各別の１つのドキュメントを分析することによって、前記対応の第２の論理形式を生成する（６４０）ステップを含む、請求項１７に記載の方法。

【請求項１９】　前記関数は、前記ドキュメントの１つのために、クエリに関連した前記第１の論理形式と前記１つのドキュメントに関連した前記第２の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、前記ランク付けするステップは、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定するステップを含む、請求項１８に記載の方法。

【請求項２０】　クエリに関連した前記第１の論理形式の任意のものと出力ドキュメント集合内の任意のドキュメントに関連した前記第２の論理形式の任意のものとの間の前記一致は同一の一致である、請求項１９に記載の方法。

【請求項２１】　システムはクライエントコンピュータをさらに含み、前記方法はクライエントコンピュータにおいて、

ユーザからのクエリを取得する（６０５；１２０５）ステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示する（６７０；１２７０）ステップとを含み、

システムはネットワーク接続（２０５，２１０，２１５）を介してクライエントコンピュータに接続されるサーバ（２２０）をさらに含み、前記方法はサーバにおいて、

クライエントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエントコンピュータに与えるステップとを含む、請求項１９に記載の方法。

【請求項２２】　前記第１の論理形式および前記第２の論理形式の各々は１つ以上の論理形式三つ組の対応の第１のリストおよび第２のリストを含み、前記第１のリスト内の前記論理形式三つ組と前記第２のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各１つの句内の、対応の論理形式グラフにおいて意味的に関係した２つの語の各々の語幹の形と、２つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項１９または２０に記載の方法。

【請求項２３】　前記１つのドキュメントのためのスコアはまた、前記１つのドキュメントのための第２の論理形式内のノード語、前記１つのドキュメント内の前記ノード語の頻度または意味的内容、前記１つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記１つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記１つのドキュメントの長さの、予め定められた関数である、請求項１９または２０に記載の方法。

【請求項２４】　前記関数は、クエリに関連した論理形式三つ組の少なくとも１つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的関係のタイプによって定義される、請求項２２に記載の方法。

【請求項２５】　前記ランク付けするステップは、

クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任

意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定するステップと、

関連した少なくとも１つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各1つのために、前記一致する論理形式三つ組の各々に関連した意味的関係によって予め規定される重み数値を用いて前記各1つのドキュメント内の一致する論理形式三つ組に重み付けして、前記１つのドキュメントのための１つ以上の重みを形成するステップと、

前記１つ以上の重みの関数として前記１つのドキュメントのためのスコアを計算するステップと、

前記ドキュメントの各１つをその前記スコアに従ってランク付けしてランク順を規定するステップとを含む、請求項２４に記載の方法。

【請求項２６】　クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの１つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項２２に記載の方法。

【請求項２７】　クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項２２に記載の方法。

【請求項２８】　サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各１つのために、リポジトリ（１０）から記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各１つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各１つにそのための関連するサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項１９、２１または２２に記載の方法。

【請求項２９】　コンピュータで実行可能な命令を記憶し、請求項１５に記載のステップを実行するためのコンピュータ読出可能媒体。

【手続補正２】
【補正対象書類名】要約書
【補正対象項目名】全文
【補正方法】変更
【補正内容】
【要約】
全体の精度を高めるために、たとえば従来の統計に基づくサーチエンジンのような情報検索エンジン（２０）によって検索された結果を処理するために自然言語処理を利用する情報検索システム（５）のための装置およびそれに付随する方法を提供する。具体的には、このようなサーチは最終的に検索されたドキュメントの集合を生む。このような各ドキュメントは次に自然言語処理を受けて論理形式の集合を生じる。このような各論理形式は句内の語間の意味的関係、特に主題と修飾語句との構造を語ー関係子ー語の態様で符号化する。ユーザが与えるクエリも同様に分析されてそのための対応の論理形式の集合を生み出す。ドキュメントはドキュメントおよびクエリからの論理形式の予め規定された関数としてランク付けされる。具体的には、クエリのための論理形式の集合は、検索されたドキュメントの各々のための論理形式の集合と比較されて両方の集合内のこのような任意の論理形式間の一致を確認する。少なくとも１つの一致する論理形式を有する各ドキュメントがヒューリスティックにスコア付けされ、一致する論理形式のための異なる各関係が異なる対応の予め定められた重みを割当てられる。このような各ドキュメントのスコアはたとえば、その独自に一致する論理形式の重みの予め規定された関数である。最後に、保持されたドキュメントがスコアの高い順にランク付けされてその順でユーザに提示される。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

Interr tal Application No
PCT/US 98/09711

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category '	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 92 04681 A (GTE LABORATORIES INC) 19 March 1992 see abstract see page 1, line 1 - page 4, line 8	1,63
A	EP 0 386 825 A (BSO BURO VOOR SYSTEEMONTWIKKEL) 12 September 1990 see abstract	1,63
A	WO 96 23265 A (BRITISH TELECOMM ;DAVIES NICHOLAS JOHN (GB); WEEKS RICHARD (GB)) 1 August 1996 see abstract	1,63
A	WO 95 29452 A (APPLE COMPUTER ;ROSE DANIEL E (US); BORNSTEIN JEREMY J (US); TIENE) 2 November 1995 see abstract	1,63

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

' Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 September 1998	16/09/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Katerbau, R

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Intern al Application No
PCT/US 98/09711

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9204681 A	19-03-1992	US 5321833 A	14-06-1994
		CA 2071485 A	01-03-1992
		EP 0497960 A	12-08-1992
		JP 5502533 T	28-04-1993
EP 0386825 A	12-09-1990	NL 8900587 A	01-10-1990
		CA 2011411 A	10-09-1990
		JP 3087975 A	12-04-1991
		US 5128865 A	07-07-1992
WO 9623265 A	01-08-1996	AU 4454996 A	14-08-1996
		BR 9606931 A	11-11-1997
		CA 2210581 A	01-08-1996
		CN 1169195 A	31-12-1997
		EP 0807291 A	19-11-1997
		FI 973080 A	22-07-1997
		NO 973372 A	22-09-1997
WO 9529452 A	02-11-1995	US 5724567 A	03-03-1998
		AU 2363895 A	16-11-1995

フロントページの続き

(72)発明者　コーストン，シモン・エイチ
アメリカ合衆国、98102　ワシントン州、シアトル、ボイルストン・アベニュ・イー、605、ナンバー・109
(72)発明者　ドラン，ウィリアム・ビィ
アメリカ合衆国、98052　ワシントン州、レッドモンド、ワンハンドレッドアンドフィフティサード・コート・エヌ・イー、7412
(72)発明者　バンダーウェンデ，ルーシー・エイチ
アメリカ合衆国、98008　ワシントン州、ベレビュ、エヌ・イー・サーティス・ストリート、16415

Ｆターム(参考)　5B075 KK07 PP24

【要約の続き】

応の予め定められた重みを割当てられる。このような各ドキュメントのスコアはたとえば、その独自に一致する論理形式の重みの予め規定された関数である。最後に、保持されたドキュメントがスコアの高い順にランク付けされてその順でユーザに提示される。